

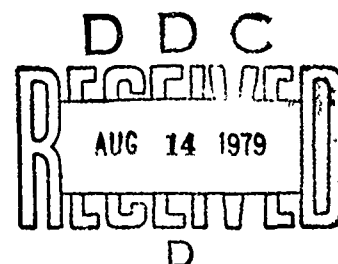
AD A072759

LEVEL II **@**
Proceedings
of
The Pacific Conference
on
Operations Research

April 23 – 28, 1979, Seoul, Korea

VOL. I

DDC FILE COPY



DISTRIBUTION STATEMENT A

Approved for public release;
Distribution Unlimited

PUBLISHED BY

Military Operations Research Society of Korea
Korean Operations Research Society

UNCLASSIFIED

SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)

REPORT DOCUMENTATION PAGE		READ INSTRUCTIONS BEFORE COMPLETING FORM
1. REPORT NUMBER 6	2. GOVT ACCESSION NO.	3. RECIPIENT'S CATALOG NUMBER <i>Final report</i>
4. TITLE (and Subtitle) Proceedings of the Pacific Conference on Operations Research, April 23-28, 1979, Seoul, Korea, <i>Volume I</i>		5. TYPE OF REPORT & PERIOD COVERED Final, 23-28 Apr 79
6. PERFORMING ORG. REPORT NUMBER		7. CONTRACT OR GRANT NUMBER(s)
8. AUTHOR(s) <i>Dr. Rak To Song, Prof Soondal Park, and Dr. Ui Chong Choe (Editors)</i>		9. PROGRAM ELEMENT, PROJECT, TASK AREA & WORK UNIT NUMBERS
9. PERFORMING ORGANIZATION NAME AND ADDRESS Military Operations Research Society of Korea, Seoul. Korean Operations Research Society		10. REPORT DATE 10 Apr 79
11. CONTROLLING OFFICE NAME AND ADDRESS Korea Institute for Defense Analyses C.P.O. Box 3089 Seoul, Republic of Korea		11. NUMBER OF PAGES 638
14. MONITORING AGENCY NAME & ADDRESS (if different from Controlling Office) <i>15634p1</i>		15. SECURITY CLASS. (of this report) UNCLASSIFIED
16. DISTRIBUTION STATEMENT (of this Report) Approved for public release; distribution unlimited.		15a. DECLASSIFICATION/DOWNGRADING SCHEDULE
17. DISTRIBUTION STATEMENT (of the abstract entered in Block 20, if different from Report)		
18. SUPPLEMENTARY NOTES See also VOL II and the Addendum to the Proceedings		
19. KEY WORDS (Continue on reverse side if necessary and identify by block number) Operations Research Pacific Area Cost Effectiveness Systems Analysis Korea Mathematical Modelling Conference Weapons Systems Evaluation Symposium Cost and Operational Effectiveness		
20. ABSTRACT (Continue on reverse side if necessary and identify by block number) These Proceedings and the Addendum contain the texts of the keynote speeches, the state-of-the-art lectures, and the papers presented at the Pacific Conference on Operations Research, held April 23-28, 1979, in Seoul, Korea. Analysts and other professionals representing government agencies, academic institutions and industry from 18 countries participated, and a total of 71 presentations were published in the two volumes of the Proceedings and the Addendum.		

DD FORM 1 JAN 73 1473

EDITION OF 1 NOV 65 IS OBSOLETE

UNCLASSIFIED

SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)

LEVEL

2

PROCEEDINGS
OF
THE PACIFIC CONFERENCE
ON
OPERATIONS RESEARCH
VOL. I

APRIL 23-28, 1979

SEOUL, KOREA

Accession For	
NPIS GRW&I	<input checked="checked" type="checkbox"/>
DDC TAB	<input type="checkbox"/>
Unannounced	<input type="checkbox"/>
Justification	
By _____	
Distribution/	
Availability Codes	
Dist.	Avail and/or special
A	

EDITED BY

DR. RAK TO SONG
PROF. SOONDAL PARK
DR. UI CHONG CHOE

DDC
RECEIVED
AUG 14 1979
D

DISTRIBUTION STATEMENT A

Approved for public release;
Distribution Unlimited.

70 00 12 006

This Conference was

Sponsored by

The Military Operations Research Society of Korea

The Korean Operations Research Society

in Collaboration with

The International Federation of Operational Research Societies

ORGANIZATION OF THE CONFERENCE

Chairman:

Dr. Moon Taik Shim
President, Military Operations Research Society of Korea

Co-Chairman:

Mr. Eung Kyun Shin
President, Korean Operations Research Society

Chairman of the Organizing Committee:

Maj. Gen. Jang-Nai Sohn
Vice President, Military Operations Research Society of
Korea

Vice Chairmen:

Dr. Sung Jin Kim, Korea Institute for Defense Analyses

Brig. Gen. Yon Sik Choi
Secretary General, Military Operations Research Society of
Korea

Secretary General:

Dr. Rak To Song, Korea Institute for Defense Analyses

Planning Committee:

Dr. Man Suk Song,
Mr. Hee-Myon Kwon, and
Mr. Kil Ho Chung, Korea Institute for Defense Analyses
Prof. Hyung Jae Oh, City University of Seoul

Editorial Committee:

Dr. Soon Dal Park, Seoul National University
Dr. Ui Chong Choe, Republic of Korea Navy

Publicity Committee:

LTC In Soo Kang, Republic of Korea Navy

Protocol Committee:

Mr. Kyu Pok Lee, Agency for Defense Development

Budget Committee:

Mr. Yang Il Sin, Agency for Defense Development

PREFACE

The Pacific Conference on Operations Research, held from 23 to 28 April, 1979, was sponsored by the Military Operations Research Society of Korea and the Korean Operations Research Society in collaboration with the International Federation of Operational Research Societies. These Proceedings contain the texts of the state-of-the-art lectures and papers presented at the Conference. We are most grateful to the authors and to the session chairmen for their efforts in bringing about the success of the Conference, and hope that the excellent papers published in these proceedings further stimulate the international exchange of ideas, to the benefit of all participants.

We gratefully acknowledge the financial support provided by the Ministry of National Defense, Republic of Korea, and the cooperation of the International Federation of Operational Research Societies. We are also indebted to Admiral Carlisle Trost, Mr. John Gratwick, and Dr. Joseph Sperrazza for their stimulating keynote addresses, and to Prof. B. Hutchinson, Dr. D. Schrady, Mr. Charles Wolf, Prof. S.M. Lee, Prof. T. Nishida, Dr. D. Hirshfeld, and Dr. Kong-Kyun Ro for their excellent state-of-the-art lectures. Last but not least, special thanks are due to organizing committee members Dr. Rak To Song, Dr. Soondal Park, Dr. Ui Chong Choe, Dr. Man Suk Song, Mr. Hee Myon Kwon, Mr. Kil Ho Chung, Prof. Hyung Jae Oh, Mr. In Soo Kang, and Mr. Juri Toomepuu, whose selfless contributions to the success of the Conference are immeasurable.

Moon Taik Shim, Ph.D.
Chairman

April 1979
Seoul, Korea

CONTENTS

CONTENTS OF VOL. I

The State-of-the-Art Lecture

1. Realism in Operations Research. 1
Joseph Sperrazza
2. Recent Developments and Future Directions in
Transport Systems Analysis. 12
B. Hutchinson
3. The Practice of Military Operations Research. . . 40
David A. Schrad
4. Resource Allocation and Defense Planning in
Retrospect and Prospect 52
Charles Wolf, Jr.
5. The Challenge of Operations Research in the
Developing Country 63
Sang Moon Lee
6. Some New Models of Queuing theory 80
Toshio Nishida
7. Methodological and Modelling Approaches for
Projecting Health Manpower Requirements and
Supply. 104
Kong-Kyun Ro

Session I : Transportation Systems Analysis

1. Man/Computer Interactive Technique in
Transportation Scheduling 137
Paul Tuan
2. An Application of Branch and Bound Method to
Optimize Interdependent Public Transit Network. . 153
In Won Lee
3. Solving a Distribution Problem with Dantzig-
Wolfe Decomposition: A Case Study 190
Ludo Gelders and Tony Roy

4. The Development of Fertilizer Distribution System: An Application of the Transportation Linear Programming Model 203
Yong Woon Yoon

5. An Alternative Zero-One Optimization Model for the Location of Fire Stations. 240
Dirk Oudheusden and F. Plastria

Session II : Military Operations Research (I)

1. Using the Transportation Method to Allocate Combat Aircraft Sorties in a Hostile Environment. 249
Bruce Ellwell

2. Optimal Allocation Strategies for Heterogeneous Force Differential Combat. 260
Hyung Kang Shin and Gil Chang Kim

3. Some Experiments in Search Theory. 280
Alan Washburn

4. Parametric Analysis of Main Battle Tank Mobility in Korean Terrain 287
Alan Thomas, W. Niemeyer, and R. Thibodeau

5. Fire Control System Performance Degradation When a Tank Gun Engages a Maneuvering Threat. 341
John McCarthy and H. Burke

Session III : Resource Allocation and Defense Planning

1. Naval Force Structure Planning 370
Leonard Gollobin

2. A Logistics Requirments Study. 391
Ronald Rush

3. Methodology for Assessing the True Worth of Perfect Forecasts. 405
Graham Winch

4. A Study on Performance Evaluation for a Computer System through Simulation. 442
Louis Chow and C. Chuang

5. SMART-Scientific Management Analysis and Review
Techniques for Local Financial Institutions. . . 457
Masayuki Akiyama
6. MINQUE Applied to Regression Analysis. 473
Moon Yul Ihuh

Session IV : Military Operations Research (II)

1. The Ammunition Stockpile Reliability Program . . 485
Alan Thomas and R. Eissner
2. Artillery Force Simulation Model (AFSM). 496
Alan Thomas and R. Sandmeyer
3. The AMSWAG Limited Visibility Study (LVS). . . . 530
John McCarthy and F. Campbell
4. Joint Munitions Effectiveness Manual and
Applications of Data 546
John McCarthy
5. Training of Personnel in the Nigerian Army
Signal Training School 571
Taiwo Abodunde and O. Fayomi
6. A Decision-Theoretic Approach to Evaluating
Effectiveness of Reconnaissance Systems in a
Target Acquisition Role. 580
Joel Hassell
7. Multidimensional Parametric Analysis Using
Response Surface Methodology and Mathematical
Programming as Applied to Military Problems. . . 592
Palmer Smith and J. Mellichamp

CONTENTS

CONTENTS OF VOL. II

Session V : Business Management (I)

1. Simulating Structured Scenarios for Corporate Planning. 616
John Tydeman and R. Mitchell
2. Comparison of Lot-Sizing Techniques for Multi-Level Material Requirements Planning Systems 640
Sung Hyon Park and Eun H. Park
3. A General Purpose Simulation System and Its Application within a Manufacturing Company. . . 662
Grahame D. Craig
4. Optimal Production Planning for a Pencil Industry: A Case Study. 702
Voratas Kachitvichyanukul and N. Sharif
5. Manpower Management System in Design Office . . 716
William Shei
6. The Optimal Inventory System of a Jute Mill . . 726
Pakorn Adulbhan and S. Svetasreni
7. Application of Models in Estimating the Quality Cost Function 747
Jens Dahlgaard

Session VI : Theory of Operations Research

1. Generating Correlated Random Variables. 779
Kim Andersen
2. A Technique for Solving Linear Programming Problems with Multiple Objectives 801
Marion Tabucanon and P. Adulbhan
3. Maximum Flows in a Vertex Weighted Network. . . 812
Gin-Hor Chan

4. Decentralized Approach to AGC of Two-Area Interconnected Power System	825
<i>Suvalai Glankwamdee</i>	
5. Deformation Method Using Parametric Approach for Solving Nonlinear Programming Problems. . .	831
<i>Yong Joon Ryang and H. Mine</i>	
6. A Corollary of the Laplace Transform Convolution Theorem and Its Application to Asymptotic Distributions	848
<i>Chang Sup Sung</i>	
7. Liapunov Technique for Nonlinear Programming	861
<i>P. Rao and P. Janakiraman</i>	
8. A Signal Flow Graph Method of Goal Programming Model.	871
<i>Chul Soo Lee</i>	

Session VII : Military Operations Research (III)

1. Management Issues and Decisions in Armor COEAs and Studies.	881
<i>Thomas Cavin</i>	
2. Cost Effectiveness Study of the Army's Training Extension Course (TEC) Program (ATECP)	920
<i>Thurston Pike</i>	
3. Korean Armor/Anti-Armor Analysis and a Comparison of Battle Analysis Methodologies	940
<i>Bernard Dunn, J. Toomey, and Eun Sang Won</i>	
4. Combat Sample Generator.	959
<i>Odie B. Richardson</i>	
5. An Application of Network Simulation to Force Analysis	969
<i>Richard O. Nugent</i>	
6. Combat Effectiveness, COEAs and Training Analysis	1002
<i>Lindsay Phillips</i>	

7. Resource Allocations in Force-On-Force/Force
Mix Analysis. 1016
Robert Hunt, P. Billingsley, and C. Lail
8. Politico-Military Simulation Methodology. 1034
James Motley

Session VIII : Business Management (II)

1. An Application of Stochastic Dynamic Programming
for Determining the Optimal Repair Limits for
a Fleet of Vehicles 1042
Himangshu Paul
2. An Algorithm for Inefficient Repair Policies. 1055
Tapas Sarkar
3. Optimum Inspection-Ordering Policies with Two
Types of Lead Times 1084
Naoto Kcio and S. Osaki
4. A Chance-Constrained Zero-One Programming Model
for Resource Constrained Capital Budgeting. 1095
B. N. Lohani
5. Loan Portfolio Analysis Under Probability
Criterium of a Typical Agricultural Credit
Institution in the Philippines: A Case Study. 1103
Victor Tan and H. Mandig
6. On a Generalized Ordering Policy. 1148
Shunji Osaki
7. Capacity Installation of Two Related
Equipment with Conversion Possibility 1157
Chan Onn Fong

Session IX : Public Administration

1. Norenwable Energy Consumption Forecasting by
Growth Curves 1177
M. Sharif, S. Khan, and M. Islam

2. The Forecasting Model for Korea Electric Power Consumption.	1199
<i>Sukho Kang</i>	
3. On Energy System Modelling for Policy Analysis	1210
<i>Byong Hun Ahn</i>	
4. The Capacitated p-Median Model for the Location of Facilities in the Public Sector	1236
<i>L. Kaufman and F. Broeckx</i>	
5. Operations Research Applications to Tourism.	1250
<i>Turgut Var, W. Swart, and C. Gearing</i>	
6. Multicriteria Decomposition in A Decentralized Organization	1272
<i>Bu Ho Roh and Sang Moon Lee</i>	
7. Application of the Hierarchical Planning Approach to Regional Development and Management in the Philippines.	1286
<i>Jona Bargur and F. Mero</i>	

VOL. I

REALISM IN MILITARY OPERATIONS RESEARCH

Dr. Joseph Sperrazza
Director

US Army Materiel Systems Analysis Activity
(USAMSAA)
Aberdeen Proving Ground, MD 21005, U.S.A.

1. INTRODUCTION

REALISM in any endeavor is a critical input. In our profession, MILITARY OPERATIONS RESEARCH, it is the most critical input. I want to emphasize, unequivocally, the requirement for realistic input data. During these sessions, you will hear much about operations research and system analysis techniques, but I suspect there will be very little on data collection, input data and validation of results. Without these elements, OR/SA may be a dangerous tool and can promote many poor decisions. The requirement for realistic input data is paramount to the success of the OR/SA community.

Let Us Look at the History of Our Profession. In the beginning OR/SA was called operational research. Initially, British operations research in World War II was on radiolocation. The early OR community in Britain had a distinct advantage over the OR community of today. They had immediate feedback on their analyses, and the opportunity to compare analytic results with actual real world results. They did not have the luxury of sophisticated methods and computer techniques available today, but they made extensive use of real world data in their analyses. Are we doing that today? I am convinced that the present OR/SA community is overcome by the sophisticated methods and computer techniques available today and doesn't place enough emphasis on the use of real world data.

2. MILITARY REQUIREMENTS

Why Do I Feel This Lack of Realism? Let me discuss several examples that I feel illustrate the necessity to return to the data collection, review of input and validation phases of OR. The first area concerns OPERATIONAL REQUIREMENTS. In the U.S., as in most of our countries, the development programs are based on operational requirements formulated by the Military branches.

An example of a critical requirement is mean time between failure (MTBF) for the system. When we applied OR/SA techniques to three different systems under development (a radar, an engine, and a helicopter) we soon recognized that from an operational point of view the original requirements were much too stringent.

TABLE 1. MEAN TIME BETWEEN FAILURE REQUIREMENTS

<u>SYSTEM</u>	<u>MTBF (HOURS)</u>
Air Defense Radar	600
Engine	1200
Helicopter	75

In the case of the air defense radar, the original MTBF was established without analyzing the operation mission and the realistic engagement requirements. An effort was undertaken to analyze the probability of the system being available as a function of attack time of enemy aircraft and the system MTBF. Given a failure occurred, various repair times were also analyzed. Figure 1 shows that the original MTBF requirement of 600 hours was not required.

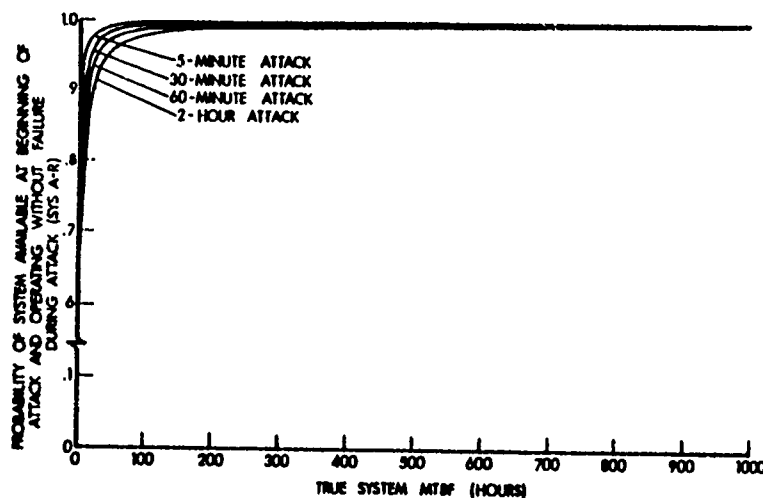


Figure 1. Effect on Air Defense Radar Availability-Reliability of Varying True System MTBF for Various Attack Times (30-Minute Repair Time).

The requirement stated for the engine directly impacted on the expected MTBF for the helicopter. When test results indicated that the engine could not meet the 1200 hour requirement, the logical question was how would a change in the engine requirement influence the availability of the aircraft for realistically defined missions.

AMSAA initiated an effort to analyze the mission of the helicopter. Based on analyses of many mission scenarios, it was determined that the maximum duration of any mission was about 3 hours. Figure 2 shows the probability that an aircraft is available and successfully completes its mission as a function of the MTBF of the aircraft. Significant for the one and three-hour missions is that a reduction of the original 75-hour MTBF for the helicopter is not very critical.

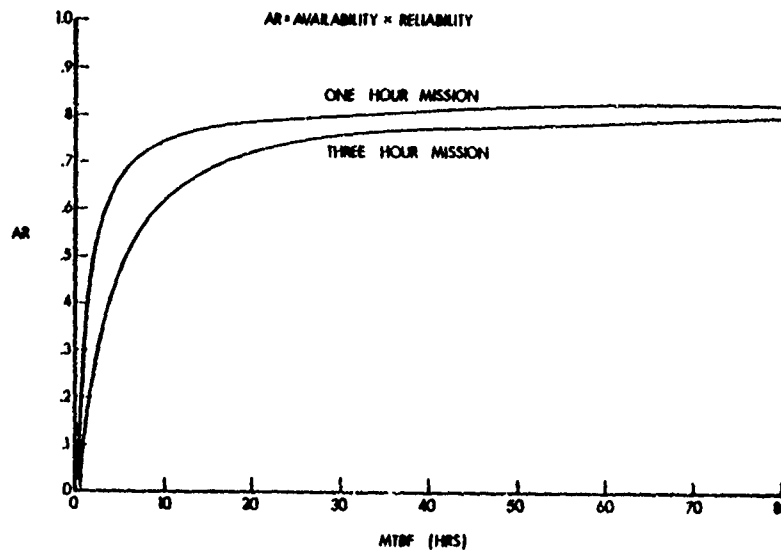


Figure 2. Probability Helicopter is Available and Successfully Completes Mission.

The interrelationship of the MTBF of the engine and helicopter was further analyzed since the engine was not meeting the requirement. Data from the testing of the helicopter system indicated that the MTBF of the system less the engine was 86 hours. Taking this into account, the data in Figure 3 were developed. The significant point here is that the system MTBF being reduced to 50 hours, could be met if the engine were to achieve approximately 200 hours MTBF. The

test results for the engine showed it could easily meet this requirement and could be changed without impacting total system availability (Figure 2).

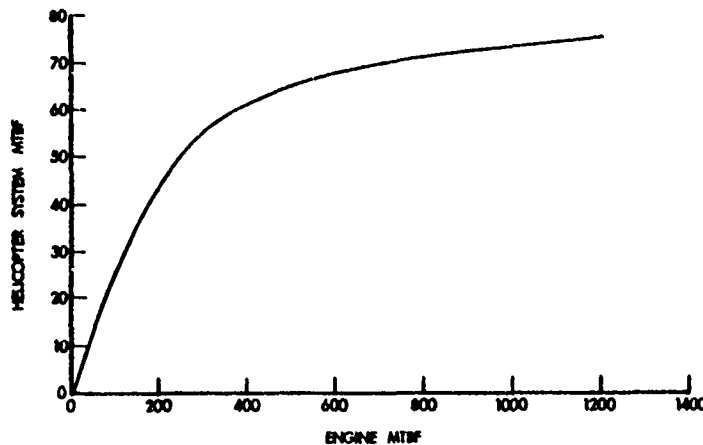


Figure 3. Helicopter MTBF less Engine MTBF = 86 hrs.

The overall result of these analyses was a change to the stated requirements as shown in Table 2, and elimination of further development and testing to meet unrealistic requirements. These are only a few examples of where military OR/SA has and can influence requirements. I urge each of you to increase your study efforts to influence this most critical phase of military development.

TABLE 2. MEAN TIME BETWEEN FAILURE REQUIREMENTS

System	MTBF (Hours)	
	Original Requirement	Modified
Air Defense Radar	600	200
Engine	1200	400
Helicopter	75	50

3. SURVIVABILITY

Another example, again with electronic equipment, where OR/SA has been applied, but unsuccessfully, concerns the survivability of shelters. We spend millions of dollars on electronic gear. Then we house them in thin walled shelters which are highly vulnerable to fragments. In the U.S., we have conducted tests and carried out extensive analyses to quantify the benefits of additional ballistic protection. We have conducted full scale verification tests to support our analytic conclusions. Figure 4 shows the improvement in survivability of a standard shelter as a function of weight of ballistic protection.

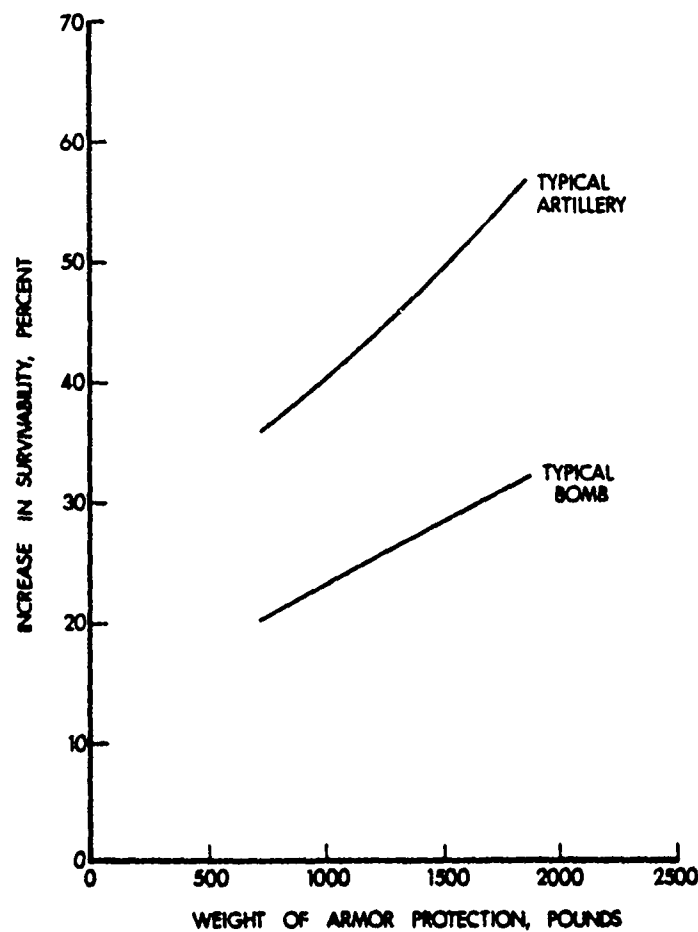


Figure 4. Survivability Increase in C³ System Hardened Shelter.

I raise this case because I feel this type application of simple, straight forward OR/SA supported by realistic data offers tremendous opportunity for future success as new equipment becomes more expensive, more complex and requires longer development cycles.

I cannot stress too strongly the need to address survivability. OR/SA can provide significant guidance and insight in this area. We can identify realistic ways to improve survivability through equipment modifications and tactical usage. I urge you to accept this challenge.

4. WEAPON DESIGN AND EVALUATION

Now let us address the area of weapon design and evaluation. In the U.S. we have two general classes of targets - air and ground targets. As one might expect, the design, development and estimated utility of weapons are very much a function of the defeat criteria. I question how realistic our kill criteria are.

In the case of air targets we have several criteria, but the two most widely used are:

A-Kill An aircraft falls out of manned control in less than 5 minutes.

K-Kill An aircraft falls out of manned control in less than 30 seconds.

These criteria were developed after World War II to allow the OR/SA people to conduct vulnerability assessments. There was little consideration given to the realistic battlefield environment, the threat and desired results. These criteria have been used for over 30 years to compare the relative effectiveness of systems.

However, when these criteria start to dictate design of weapon systems, I become alarmed. Usually we state a requirement for a K-Kill at long ranges for an air defense gun which leads to a large caliber system. It is driven by the desire to produce a K-Kill before an enemy aircraft can drop its ordnance.

An investigation of combat data indicates that aircraft attacking ground targets, need not be catastrophically destroyed. We point out that the mere presence of air defense guns, degrades delivery accuracy of air delivered ordnance and also results in aircraft damaged that do not

return. Table 3 shows the relative accuracy of air delivered ordnance experienced by the U.S. in Vietnam for no air defense, light and heavy defenses. The degraded accuracy has a significant effect on the damage level produced.

TABLE 3. ACCURACY OF AIR DELIVERED ORDNANCE

<u>Air Defense</u>	<u>Relative Accuracy</u>
None	1.0
Light	1.5
Heavy	3.1

Of equal or greater importance is the fact that aircraft that are hit, but not necessarily killed, have difficulty in completing their mission, and require repair time. In Vietnam we found that of the aircraft hit by 23mm and 37mm projectiles and not killed, over 50 percent aborted the mission. For some aircraft hit that returned to base, repair time in excess of 30 days occurred.

So, I question whether or not the K-kill criteria is realistic when a large numerical threat exists, and large caliber systems have limited rates of fire and present greater logistical burdens. Thus, to realistically determine the characteristics of a system, we must go far beyond the single statement of a kill criteria.

In the case of air-to-air missiles, again the desire for a K-kill is leading to greater sophistication in warhead design, fuzing and guidance in an environment of reduced weight for missiles.

I urge the OR/SA community not to accept blindly the criteria developed in the past, but concentrate on analyzing the real world situation and apply or develop appropriate criteria. The threat, technology and operational requirements of today are entirely different from those of WW II and must be reflected in the criteria we use for evaluation. Although I've only discussed air targets the same arguments hold for ground targets.

5. ENVIRONMENTAL CONSIDERATIONS

Another critical factor to consider is the environment. Given a weapon or weapon systems for evaluation, the environment in which it is to perform is a basic, critical consideration. Today in the U.S. we have come to depend on sophisticated and accurate systems. The OR/SA community has done many studies in support of decisions on laser guided (COPPERHEAD), Smart Bombs, electro optical (MAVERICK) and wireguided (TOW) systems.

How realistic are these analyses? Until the 1973 Arab/Israeli War, I must say they were not very realistic. The extensive use of guided systems in 1973, prompted us to relearn the lessons on the use of smoke. In the U.S. considerable effort is being expanded to quantify the effects of smoke, dust, and poor weather. I hope we have not gone too far on precision systems!

Going back to the Vietnam War, again reminds me of how the OR/SA community lacked realism in its treatment of the environment. We traditionally evaluated weapons in open terrain, however, in Vietnam we had marsh grass and rain forests.

The OR/SA community had generated data recommending usage of weapons and fuzes based on the open terrain. Needless to say, the data base did not apply and as a result we launched a very aggressive data collection program.

Today we have a much better data base for realistic assessment of weapons. But as weapon designs change and become more sophisticated, these data base must be updated. Otherwise, OR/SA in the weapons area could mislead the decision maker.

Before leaving this area, I raise a word of caution. In accepting data on new systems or analyses on systems, we must take into account the impact of the particular environment considered and the appropriate operational application.

6. LOGISTICS SUPPORT

Now, after a system is fielded - how about its logistical support? The most effective system is useless in the field if it cannot be supported. The logistical support system is very complex and costly. OR/SA can definitely make a significant contribution in this area, however, to date very limited application of OR/SA has been seen.

In the U.S. considerable emphasis is being placed on spare parts provisioning - for both nominal wearout and combat damage. Again, a review of Vietnam combat data has led us to relearn the lesson that the spare parts required for combat are much different than those required in peacetime operations.

This is illustrated very clearly by examining damage to armored vehicles. When an armored vehicle is damaged by a threat weapon, the wiring harness suffers damage. We do not stock wiring harness because they do not wearout in peacetime. Another example is the fuel cell in helicopters. The fuel cell represents a large presented area on the helicopter and is frequently damaged in combat. In the U.S. we stock very few fuel cells - again because they do not wearout.

We have not performed too well in the logistical area. A significant study effort is required to address realistically the required logistical support for key systems. In addition, reliable data collection efforts are required. To improve the data base, special sample data collection programs are established. Also, AMSAA has a unique mission of sending teams to the field to investigate the performance of equipment and as a result identifies problem areas for future investigation. These investigations range from redesign to maintenance of equipment as well as identifying requirements for changes in TO&E*, training and support. There is a great deal of money, time and resources devoted to new developments, new technology - but not that much money and resources dedicated toward improving fielded equipment. Why not? That is what my field liaison program accomplishes - IMPROVING FIELDED EQUIPMENT. This program thrives on realism. We do not rely on reports, or on surveys - we rely on "face-toface" contact with the soldier in the field. The program offers an excellent check on equipment performance in the hands of the soldier versus proving ground assessments of performance.

7. SUMMARY

I have attempted to provide you my thoughts on areas where I think OR/SA needs increased emphasis. In summary, we must concentrate on a well balanced and realistic approach supported by reasonable data from many sources. Too much emphasis on one element, too much emphasis on computer techniques, too much emphasis on complex simulations can all lead to misleading results.

* Table of Organization and Equipment

I think the OR/SA military approach will have greater realism if we remember analysis is a cyclical procedure involving predictive models, supported by data and tests. A key to success is to take advantage of field operations including combat data collection.

Let us remember OR is not Operations Research - It is Operational Research.

I wish you success in your search of methods and data to solve today's problems.

RECENT DEVELOPMENTS AND FUTURE DIRECTIONS
IN TRANSPORT SYSTEMS ANALYSIS

B.G. Hutchinson

Department of Civil Engineering
University of Waterloo
Waterloo, Ontario. N2L 3G1, Canada

ABSTRACT. The broad patterns of expenditures on transport services are outlined and the general characteristics of the typical planning process are described along with the role played by transport systems analysis. The properties of the transport systems analysis tools used in several of the major transport sectors are then described where the major emphasis is on the urban transport sector. The adequacies and deficiencies of available systems analysis tools are discussed and the opportunities for improvements are identified. It is concluded that the most potentially productive areas of research are at the interfaces between transport systems and the socio-economic environments that they serve, rather than within the transport sector itself.

1. INTRODUCTION

Large scale transport studies employing computer-based systems analysis techniques began to emerge in several rapidly growing North American urban areas during the 1950s with the seminal studies originating in Detroit [1], Chicago [2] and Toronto [3]. Since that time comprehensive transport planning studies have been conducted in hundreds of cities throughout the world [4]. The methodology established for metropolitan transport studies [5] has been adapted for use in other transport sectors such as inter-city transport with the studies performed in the Boston-New York-Washington corridor of the U.S.A. during the 1960s [6] representing one of the first major applications.

The primary aim of this paper is to provide an overview of the accomplishments of transport systems analysis during the past two decades and to identify potentially productive opportunities for improvements to existing techniques. The bulk of the observations made in this paper pertain to the urban transport sector with some comments on the inter-city and regional transport sectors. In addition, the particular thrust of the paper is conditioned by the Canadian environment and some of the observations made may not apply to other countries of the Pacific Rim because of the tremendous diversity of social, cultural and economic characteristics that exist.

2. IMPORTANCE OF THE DIFFERENT TRANSPORT SECTORS

Fig. 1 summarizes the components of transport expenditures in the U.S.A. in 1972 [7]. While some changes may have occurred during the past six years because of economic growth and change the broad patterns of revenues and expenditures would be roughly similar at the present time. The diagram illustrates that domestic expenditures represented about 97.5 percent of the total domestic plus international expenditures on transport. About 85 percent of all expenditures were on highway transport with approximately half of these expenditures on the passenger car. The bulk of these private car expenditures were for urban transport and Fig. 1 also illustrates that about 21 percent of all transport expenditures were for local truck transport in urban areas. In other words about 65 percent of the transport expenditures in the U.S.A. in 1972 were for the movement of people and goods within urban areas.

Fig. 1 shows that approximately 30 percent of all transport expenditures in the U.S.A. in 1972 were for inter-city

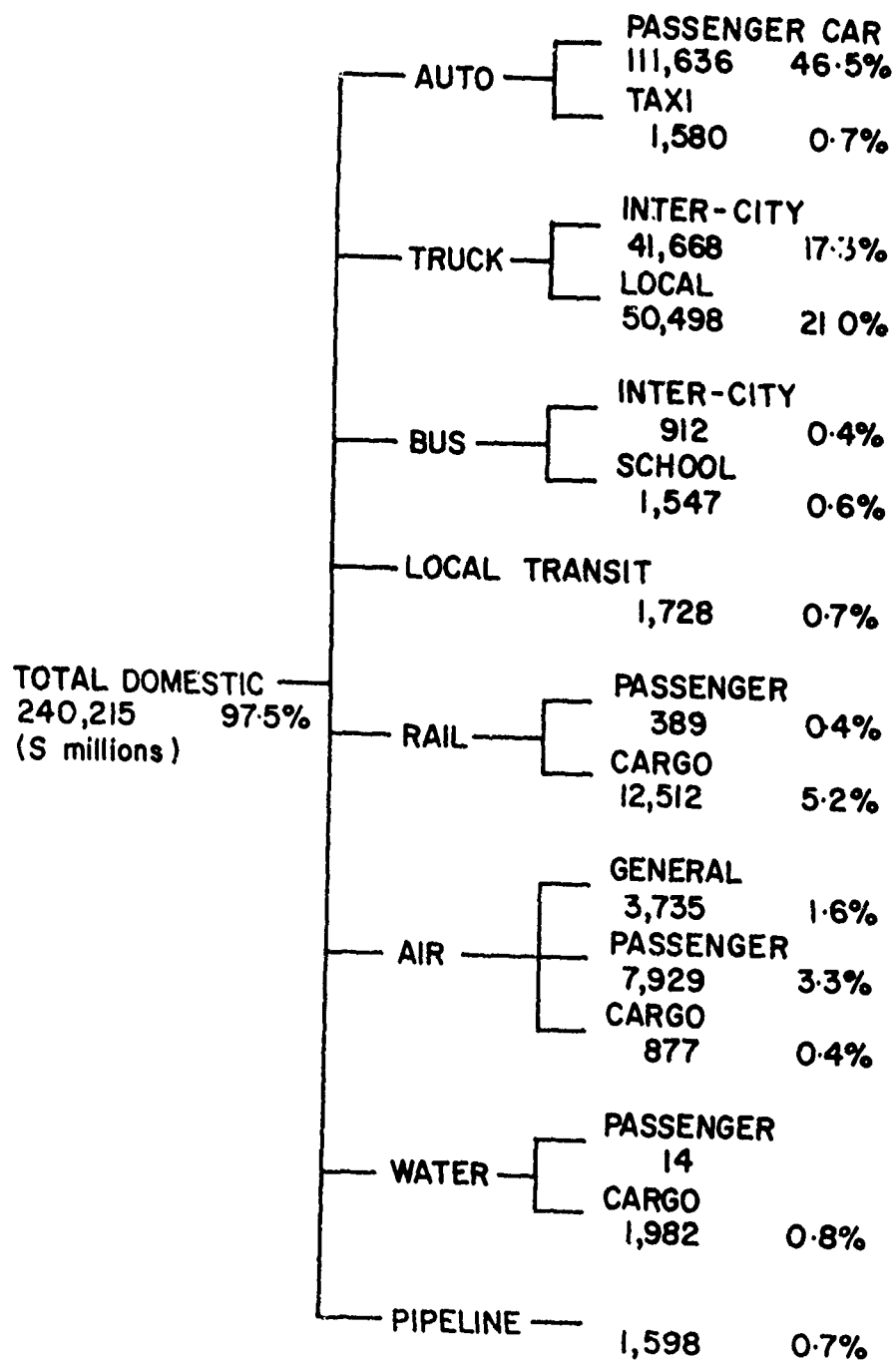


FIG. 1 - DOMESTIC TRANSPORT EXPENDITURES U.S.A., 1972

transport services. Inter-city truck movements accounted for more than half of these expenditures, with the rail transport of freight and the air transport of passengers representing the next most important modes of inter-city transport.

2.1. International Comparisons

The relative importance of the different transport sectors may vary significantly between countries depending on factors such as the degree of urbanization, the geographic scale of the country, the character of the economy and so on. One example is provided by Fig. 2 in which the rail net-tonne-kilometres and the air passenger-kilometres in 1960 and 1974 are shown for selected countries [8]. The upper diagram shows that rail freight movements in Canada in 1974 were about one-sixth of the movements in the U.S.A., while the Canadian population was only one-tenth of that of the U.S.A. Population is more dispersed in Canada than in the U.S.A. and market areas for manufactured goods are broader. It is interesting to note from Fig. 2 that rail freight movement growth almost doubled in both Canada and Australia between 1960 and 1974. Most of this growth resulted from increased exports of bulk natural resources such as coal and iron ore as well as grain shipments.

The lower part of Fig. 2 illustrates that the air travel mode is more important in North America than in Western Europe and Japan although the rapid rates of growth in air travel in both West Germany and Japan should be noted. Inter-city travel distances are very large in the U.S.A., Canada and Australia and the much higher amounts of air passenger travel per capita in these countries are illustrated in Fig. 2.

3. THE TRANSPORT PLANNING PROCESS

The broad sequence of steps involved in the typical transport planning study is illustrated in Fig. 3. Transport demand is a derived demand and Fig. 3 shows that calibrated transport demand models are used along with estimates of the future distribution of human activities to estimate future travel demands. These calibrated systems analysis models are also used to estimate the transport network equilibrium flows that are likely to result from the interaction between demand and a particular supply strategy. Transport supply strategies are not only concerned with the provision of a particular technology but also with the operating and pricing policies applied to a technology.

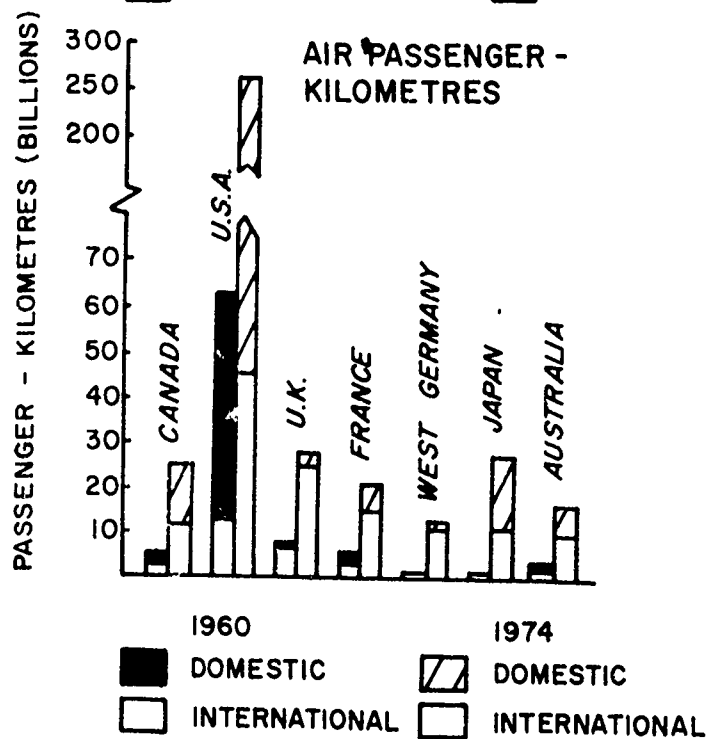
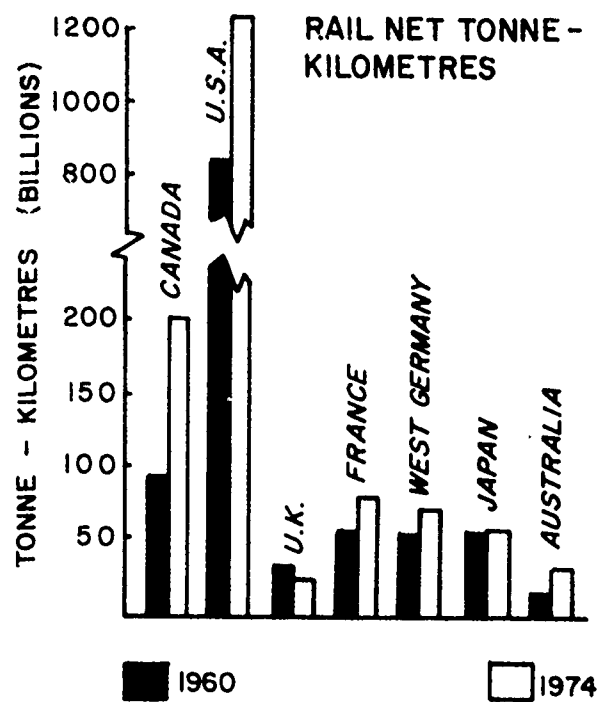


FIG. 2 - INTER-CITY TRANSPORT DEMANDS IN SELECTED COUNTRIES

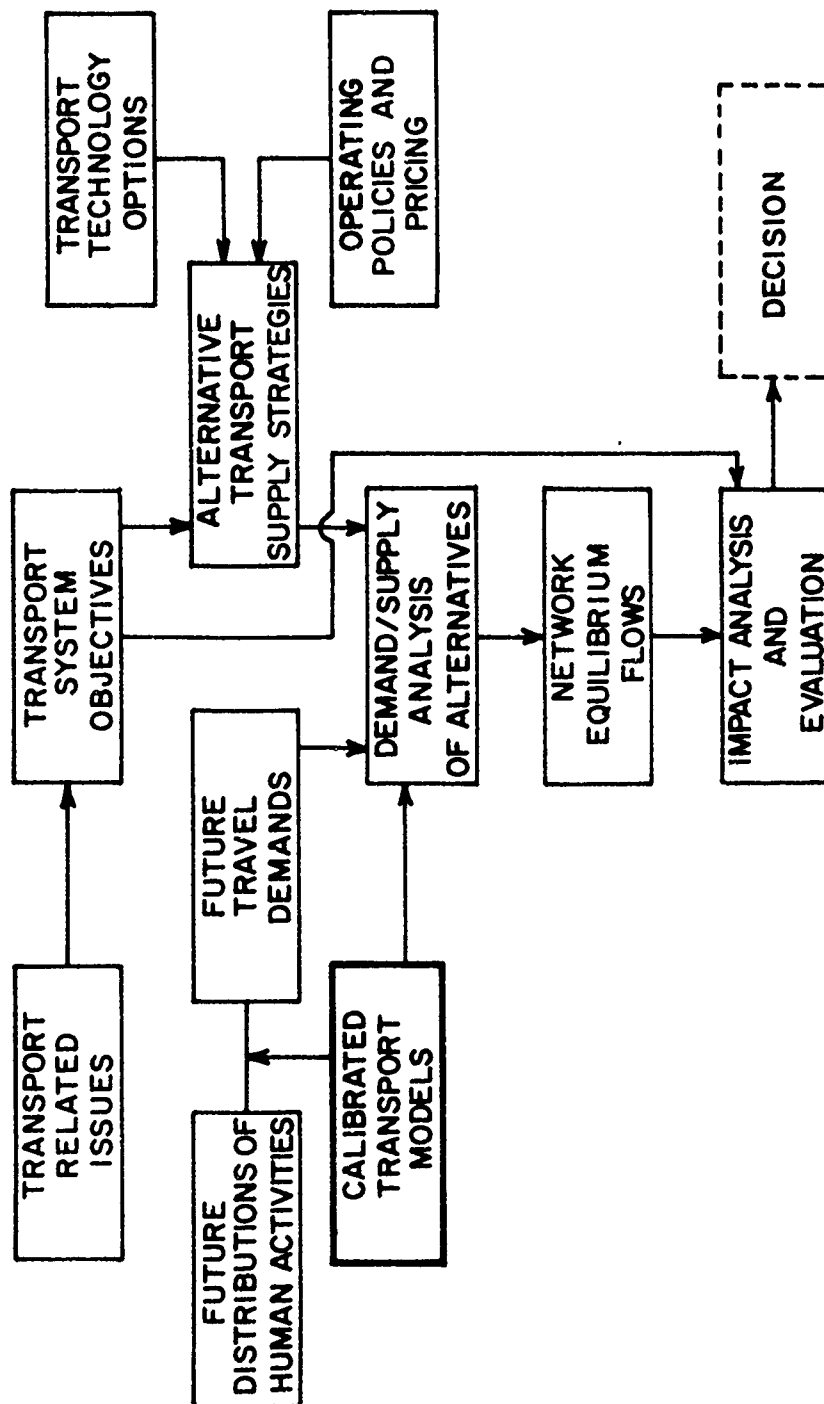


FIG. 3 - BROAD STRUCTURE OF TRANSPORT PLANNING PROCESS

The final phase of the process illustrated in Fig. 3 is concerned with the analysis of the probable impacts of alternative strategies and their overall evaluation. The types of impacts considered will vary between transport sectors but for the urban transport sector these might include noise and air pollution effects, energy consumption, changes in the mobility of the various socio-economic groups living in a community and the economic efficiencies of the proposed investments.

Fig. 3 suggests that the choice of a particular transport strategy from those considered to be technically and economically feasible is not part of the technical process of transport planning but is the prerogative of governments in most transport sectors. Technical information on the probable impacts of alternative strategies allows informed debate and the necessary arbitration by politicians to take place.

The transport planning process that has emerged from the many urban transport planning studies represents one of the first civilian applications of modern systems analysis techniques. An important principle of systems analysis is that the technical components of the process are directed towards explicitly defined objectives. Objectives evolve from the specific issues to be resolved by the planning process and the objectives have an important influence on the strategies proposed as well as the way in which these alternatives are analyzed and evaluated. Changes in the objectives set for transport systems will stimulate changes in the technical process used to achieve these new objectives.

3.1. Hierarchical Levels of Planning

Transport planning and design activities are carried on at a number of levels of detail and with respect to a number of time horizons. Three broad planning levels may be identified and these are the transport systems management level, the transport systems planning level and the strategic planning level. Transport systems management is short-run in nature and is directed towards the optimization of existing transport facilities. Transport systems planning is usually directed towards time horizons of 10 to 15 years and is concerned primarily with the identification of capital investment opportunities. Strategic transport studies are concerned with time horizons of 20 to 30 years and typically the transport system is considered as just one element of a broad development strategy. Clearly the systems analysis tools that might support each of these levels of planning will be different in character. The techniques outlined later in this paper are concerned primarily with the transport systems

planning level.

4. URBAN TRANSPORT

Much of the research and development work on transport systems analysis during the past two decades has been directed towards the urban transport sector. The dominant transport-related issues have changed very dramatically during this period and the characteristics of the analysis tools have been re-oriented to these changing circumstances.

4.1. Changing Objectives of Urban Transport Planning

In most of the metropolitan transport planning studies conducted during the late 1950s and early 1960s the planning objective was one of reducing road congestion given the conditions of sustained urban growth and rapidly increasing car ownership. The pre-occupation of urban transport planning with the development of long-range capital investment programs began to change during the late 1960s. Many communities throughout the world abandoned plans for freeways and fixed route public transport systems that had been developed in the earlier studies.

In the late 1960s and early 1970s the objectives of urban transport planning studies broadened in a rather dramatic way with mobility objectives becoming just part of a broad set of objectives concerned with environmental impacts, energy conservation and land development. These changing objectives required new policy analysis tools that were capable of testing a range of policy responses such as pricing, traffic restraint as well as longer range land development alternatives.

4.2. Estimating Urban Travel Behaviour

Urban transport systems analysis models attempt to capture the transport decision making behaviour of individuals and the ways in which this behaviour might be influenced by changes in the transport policy environment affecting individual trip makers. The trip making behaviour of individuals has been represented traditionally by a sequence of transport sub-models of the following form:

$$t_{ij}^{mr} = p_i s_{ij} s_{ij}^m s_{ij}^{mr} \quad (1)$$

where t_{ij}^{mr} = The number of trips from some zone i
to a zone j by mode m and modal
network route r

- p_i = The number of trips produced in zone i
- s_{ij} = The proportion of trips produced in zone i that travel to destinations in zone j
- s_{ij}^m = The proportion of trips between i and j that travel by mode m
- s_{ij}^{mr} = The proportion of trips between zones i and j by mode m that travel by route r

The travel demand matrices calculated by equation (1) are conditional in the sense that they are for a particular spatial distribution of human activities and a specific transport supply strategy. Equation (1) may be stratified by socio-economic group and the location of the modal split sub-model, s_{ij}^m , in the sub-model sequence varies [5].

A fundamental issue that arises in estimating the parameters of the sub-models of equation (1) is the extent to which travel decisions may be treated as a sequence of separate choices or must be treated simultaneously. Most of the operational forms of equation (1) that have been developed assume that choice decisions are separable and may be built up from a sequence of conditional probability measures which reflect the chains of decisions made by urban trip makers. For example, s_{ij}^m may be interpreted as the conditional probability that a trip maker will choose transport mode m given that the trip is between zones i and j , and so on. At each stage of decision, choice is viewed as being conditioned on fixed preceding decisions and optimal succeeding decisions. If the utility derived from each level of decision is additive then decisions may be treated as being separable. If utility is not separable then the decisions must be analyzed simultaneously. This is a very critical issue since it governs the way in which model structures are specified and estimated.

The first sub-model in equation (1) is trip generation. Trip generation has been largely a matter of empirical investigation in which the observed rates of trip generation for some system of basic spatial units have been related to measures of the amount of human activity in those spatial units through regression analysis [5]. Traditionally, trip generation rates established in this way have been assumed to be inelastic to transport supply. While this approach was satisfactory in the earlier studies concerned with estimating future car traffic volumes it is unsatisfactory for the analysis of policies geared to traffic restraint.

Perhaps the most important and intractable problem in transport systems analysis is understanding the patterns of human activity interaction. There is a need not only for understanding existing spatial linkages but how these linkages might change and develop over time in response to changes in the transport system and land development. The spatial linkages that exist in any urban area have been built up over many years and reflect a myriad of individual location decisions made by households and institutions. Most of the available spatial interaction models available for estimating the s_{ij} in equation (1) are of the comparative static type in that they are estimated from cross-sectional travel data at one point in time. Some of these models require fully specified activity distributions at both the origin and destination ends while other models require only partially specified activity distributions with the models estimating the locations of the remaining activities along with the spatial linkages. Some of the spatial interaction models have their origins in the gravity and potential concepts of social science while others are based on certain principles of mathematical programming.

4.2.1. Gravity Type Models

There are two basic components of spatial interaction and these are the number of trips between any pair of zones, t_{ij} , and the associated costs of travelling between the zones, c_{ij} . In urban transport planning travel costs are used in the generalized sense and include monetary costs, time costs and comfort/convenience costs. A third dimension of the problem is that the row and column totals of any trip interchange matrix must be equal to the so-called trip-end constraint equations:

$$\sum_j t_{ij} = p_i \quad (2)$$

= The total number of trips with origins in zone i , the so-called trip productions

$$\sum_i t_{ij} = d_j \quad (3)$$

= The total number of trips with destinations in zone j , the so-called trip attractions

Additional constraint equations may be introduced which ensure, for example, that the total amount of travel effort is equal

to some specified constant:

$$C = \sum_i \sum_j c_{ij} \quad (4)$$

Most of the spatial interaction models that have been used in transport planning studies throughout the world had their origins in the pioneering work of Voorhees [9]. Voorhees adapted some of the earlier work of social scientists who had applied simple gravity and potential concepts to the modelling of human interactions of various types including trade, migration and communications flows. The gravity type models used in most transport planning studies have been derived heuristically. In 1967 Wilson [10] made a major contribution to spatial interaction modelling by proposing a formal procedure for deriving spatial interaction models. Using some entropy maximizing concepts from statistical mechanics he showed that a family of spatial interaction models could be derived in a consistent manner from the constraints imposed on the trip matrix. The Wilson version of the production-attraction gravity model is:

$$\begin{aligned} t_{ij} &= p_i s_{ij} \\ &= p_i b_i b_j a_j e^{-\beta c_{ij}} \end{aligned} \quad (5)$$

$$b_i = [\sum_j b_j a_j e^{-\beta c_{ij}}]^{-1} \quad (6)$$

$$b_j = [\sum_i b_i p_i e^{-\beta c_{ij}}]^{-1} \quad (7)$$

The terms b_i and b_j are usually referred to as balancing factors which ensure that the constraint equations (2) and (3) are satisfied and the parameter β ensures that the constraint equation (7) is satisfied.

While gravity models have been used widely in transport planning studies at both the urban and inter-city scales their capabilities in estimating observed spatial interaction patterns have not been exhaustively examined. Hutchinson and Smith [11] have examined the extent to which state-of-the-art gravity models are capable of explaining journey to work patterns in the thirty census areas of Canada and have concluded that major improvements to the gravity model are required. Many of the residuals between the observed and estimated trip interchange flows were as large as the observed

trip interchange magnitudes. Most errors result from the attempt to calibrate a cross-sectional model to trip linkage patterns that have developed over long periods of time and where these linkage patterns reflect the spatial distributions of housing and job opportunities that existed at various points in time. It seems clear that dynamic forms of the gravity model must be developed in order to predict better the future development of trip linkages in urban areas. While Wilson [12] has proposed dynamic forms of the gravity model few serious attempts have been made to calibrate these types of models. Hutchinson and Smith [13] have described some initial attempts to calibrate improved gravity models for the fifteen census areas of Ontario.

4.2.2. Mathematical Programming Approaches

While the majority of the work on urban spatial interaction has been with gravity-type models a number of alternative approaches have been suggested [5]. One of the most interesting alternative approaches is based on the transportation problem of linear programming and Blunden and his co-workers [14, 15, 16] have formulated the trip distribution problem in the following way:

$$\text{minimize } Z = \sum_i \sum_j t_{ij} c_{ij} \quad (8)$$

$$\text{subject to } \sum_j t_{ij} = p_i \quad (9)$$

$$\sum_i t_{ij} = a_j \quad (10)$$

$$t_{ij} \geq 0, p_i \geq 0, a_j \geq 0 \quad (11)$$

The assumption of this approach which is embodied in equation (8) is that the equilibrium trip distribution state for a given human activity allocation and a particular transport network is given by the set of trip distributions which minimizes the total travel costs of all trip makers. Evans [17] has shown that the trip matrices estimated by the linear programming solution are approached by those estimated by a gravity model with β tending to infinity.

The validity of the linear programming approach rests on the assumption that residential and workplace locations, for example, are selected jointly by all locators so as to minimize collectively the total travel costs. Comparisons of trip matrices estimated by linear programming with observed

matrices show that the estimated mean trip lengths are shorter than the observed. Trip matrices estimated by linear programming assume implicitly that location decisions will be made on the basis of the marginal social costs of travel. However, locators perceive the average costs of travel in making location decisions which are lower than marginal costs and longer trip lengths result.

While the linear programming approach has deficiencies it does provide important insights into the interactions between land use and transport through the dual formulation of the problem stated in equations (8) through (11). The dual variables calculated by this formulation provide some interesting information about the travel cost implications of unit changes in the production and attraction trip ends brought about by the relocation of residences and workplaces. Information of this type is particularly useful for strategic planning where the transport implications of alternative arrangements of land use are being explored. Blunden and Black [16] provide an interesting application of the primal/dual formulation to the Sydney, Australia region.

4.2.3. Understanding Transport Mode Choice

In the earlier transport studies empirical relationships were developed between the modal choice probabilities of different person types and some very coarse indicators of transport system properties. These earlier methods were satisfactory for estimating the broad proportions of trips by public transport and private vehicles for an unchanged policy environment but they were found to be unsatisfactory for testing fare changes, the impacts of traffic restraint schemes, and so on.

Two of the principal contributions to the better understanding of modal choice decisions during the past decade have been the development of the so-called disaggregate models of transport demand and the concept of generalized transport cost. Disaggregate models of transport demand began to emerge in the early 1960s [18] and became quite well developed during the late 1960s and early 1970s [19, 20, 21]. The term disaggregate was used to reflect the fact that the analytical techniques focussed on the behaviour of individuals rather than on the average behaviour of groups of people. The basic form of disaggregate modal split models is:

$$s_{ij}^{km} = f(w^k, x_{ij}^m) / \sum_m f(w^k, x_{ij}^m) \quad (12)$$

where s_{ij}^{km} = The probability of an individual of type k choosing a transport mode m for a trip between zones i and j

$f(w^k, x_{ij}^m)$ = A function of a set of variables w^k which describe the characteristics of an individual of type k and variables x_{ij}^m which describe the characteristics of a transport mode m between zones i and j

A number of mechanisms of choice have been postulated in order to derive the character of the function used in equation (12) where these postulates have been derived from theories of choice in economics and psychology and a typical function form is the so-called logit model with the following form [22]:

$$s_{ij}^{km} = \frac{e^{-\lambda^k c_{ij}^m}}{\sum_m e^{-\lambda^k c_{ij}^m}} \quad (13)$$

where λ^k = A person-type parameter which reflects the impact that transport mode attributes have on the modal choice behaviour of person type k

c_{ij}^m = The generalized costs of using mode m between zones i and j

A critical component of the modal choice model defined in equation (13) is the generalized travel cost variable which is usually calculated from:

$$c_{ij}^m = a_1 x_{ij}^{1m} + a_2 x_{ij}^{2m} + a_3 x_{ij}^{3m} + x_j^{4m} \quad (14)$$

where x_{ij}^{1m} = The in-vehicle travel time between zones i and j by mode m

x_{ij}^{2m} = The out-of-vehicle travel time (wait, transfer) between zones i and j by mode m

x_{ij}^{3m} = The fare or other monetary costs of travel between zones i and j by mode m

x_j^{4m} = The parking charges or other terminal costs
at zone j for travel by mode m

There is still some debate about the methods of estimating the parameters of equations (13) and (14). A number of investigators have shown that the λ parameter magnitude of (13) is sensitive to the estimation method used. In addition the parameter magnitude is influenced by the structure of the generalized cost function selected.

The general form of the logit model specified in equation (13) has been used to analyze other types of transport choice situations such as decisions about car ownership, time and frequency of travel and destination choice [23].

4.2.4. Route Choice Mechanisms

The fourth and final sub-model of equation (1) is the network route choice sub-model. In many of the earlier transport studies road traffic demands were assigned to the minimum travel time path between each zonal pair on the assumption that this represented the demand for road capacity and that adequate road capacity would be provided along these paths by the construction of new facilities. Many of the transport studies performed during the 1960s recognized that road capacity could not be increased in many parts of an urban area and capacity-restrained assignments were used to estimate equilibrium traffic flows. Most of the traffic assignment procedures which have been developed are based on the assignment principle enunciated by Wardrop [24]. The basic principle proposed by Wardrop is that traffic on a network distributes itself in such a way that the travel costs on all of the routes used between any pair of origin and destination zones are equal while all unused routes have equal or greater costs.

Route assignment methods which are based on Wardrop's principle and which employ some form of capacity restraint seem to provide adequate representations of total traffic patterns [25, 26]. Florian and his co-workers [27, 28] have proposed an equilibrium assignment technique based on Wardrop's principles that provides good estimates of observed traffic flows.

4.2.5. Adequacy of Existing Models

The transport systems analysis techniques which have evolved during the past two decades tend to be rather cumber-

some, time consuming and expensive. A period of about two years is required for data collection and coding, model development and network analysis. While there have been significant improvements in each of the sub-models imbedded in equation (1) there are still significant weaknesses in the model set.

An important problem with current models is the lack of consistency both between and within sub-models. At the broadest level there is the concern about whether the choice mechanisms of equation (1) may be treated sequentially or must be examined simultaneously. Also each of the sub-models incorporates some measure of generalized travel costs and these costs are not consistent between each of the sub-models. Travel times between zones are typically used in trip distribution and route choice sub-models whereas generalized travel costs of the type specified in equation (14) have been used in many modal split models and in the evaluation of the user benefits of alternative strategies.

Perhaps the major difficulty with all of the sub-models is that while they are capable of reproducing travel behaviour observed at a particular point in time there is no guarantee that these cross-sectional type models are useful for estimating the marginal changes in future behaviour. For example, it has been noted already that gravity-type trip distribution models calibrated to cross-sectional data from one particular year simulate in some average way spatial interaction patterns that have formed over many years and stages of development but there is real concern about their ability to capture marginal changes in trip distribution patterns over time.

Models of the type defined in equation (1) produce travel demand estimates which are conditional on the exogenous specification of urban development patterns. Urban development patterns emerge simultaneously with actions in the transport sector and the sequential and conditional modelling of travel demands creates major difficulties. Modelling capabilities have been developed which attempt to capture the joint character of land development and transport.

4.3. Urban Systems Models

The transport systems analysis models discussed in the previous sections have been used principally in transport systems planning studies where the principal aim was to identify opportunities for new investments in transport capacity. More recently the techniques have been adapted for use in the shorter-run transport systems management studies.

While much of the current transport planning emphasis is on the shorter-run tactical policies there is increasing recognition of the need for longer-run strategic studies which focus on the potential for manipulating travel demands through better land development planning. Studies on this type must be supported by analytical capabilities with a very different character to those used in the shorter-run management studies and the medium-run transport systems studies.

Lowry [29] was the first to suggest linking gravity-type allocation models together in order to reduce the extent of the exogenously specified constraints imposed on the distributions of urban activities. Most modelling approaches specify part of the employment distribution exogenously with the model allocating the spatial distributions of the remaining employment and the household sector endogenously in response to specified land development and transport policies. Wilson [12] and Batty [30] have reviewed much of the work on urban activity systems models conducted over the past two decades. Fig. 4 illustrates the structure of urban systems models of the type mentioned previously. The allocation functions imbedded in this type of urban systems model are usually of the gravity type with the following household allocation function providing an example:

$$s_{ij}^{wk} = \ell_j^{hk} e^{-\beta^k c_{ij}} / \sum_j \ell_j^{hk} e^{-\beta^k c_{ij}} \quad (15)$$

where s_{ij}^{wk} = The probability of an employment of person type k working in zone i and living in zone j

ℓ_j^{hk} = The amount of land available in zone j for residential development that is compatible with person type k housing preferences

and the β^k and c_{ij} reflect the effect of the transport system properties on the residential location choice behaviour of person type k .

Hutchinson [31] has described an application of this type of model to planning problems in the Toronto, Canada region while Sarna and Hutchinson [32] have described an application to the Delhi region of India. Said and Hutchinson [33] have outlined a model framework which is directed towards estimating the time staged development of urban area.

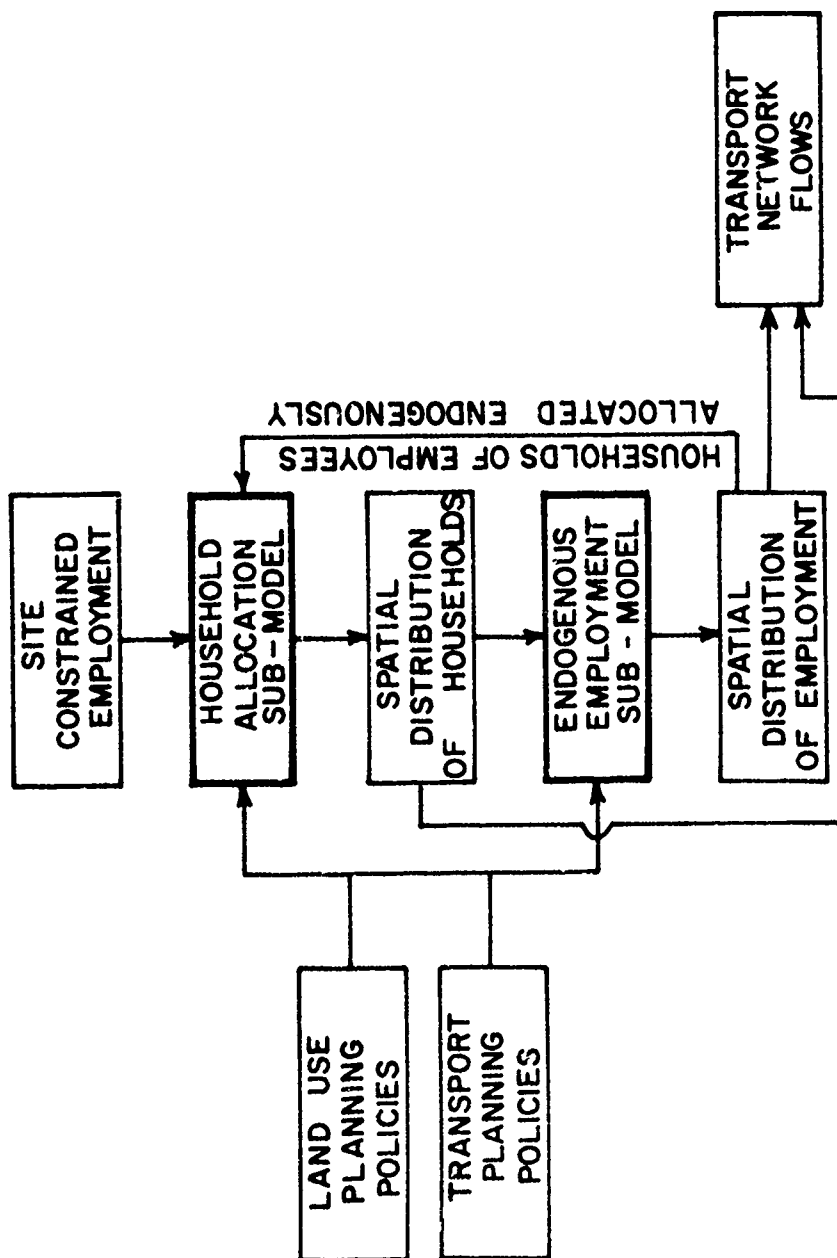


FIG 4 - BROAD STRUCTURE OF URBAN ACTIVITY SYSTEM MODEL

Normative urban systems models have been developed which search for urban development patterns that optimize a particular objective function [34, 35]. Gupta and Hutchinson [36] have described a land use-transport model for optimizing development in the Delhi region of India. The model searches for regional development configurations that minimize the combined costs of providing urban infrastructure and inter-urban transport costs. The objective function is a non-linear function of the regional activity distribution and a sequential search procedure is used to identify the least cost alternative. The search procedure operates by allocating a hypothetical increment of population growth to each urban centre in the region and then estimates the employment growth necessary to support this population increment along with the expected increase in inter-city travel demands. The total costs of development are then estimated and the population increment is allocated to the urban centre with the minimum marginal costs subject to any constraints on population and employment holding capacities and on inter-city transport capacities. The search procedure continues until the total expected regional population growth has been allocated. Simple policy analysis tools of this type may be used to highlight the costs and impacts of alternative regional development strategies.

While the behavioural and normative urban systems models which have been developed and used over the past two decades have provided useful information for planning studies they represent highly simplified abstractions of the urban development process. A large amount of information must be specified exogenously to all urban systems models and the outcomes predicted by the models are influenced strongly by these constraints. For example, the amount of vacant land available for residential development at any particular time is usually identified exogenously to most models. In the North American environment this is influenced by a myriad of decisions made by governments, land developers, planners and builders. The development of truly policy-sensitive urban systems models requires that many of these factors be incorporated in the model structure rather than being specified exogenously by the policy analyst. In addition, the locational choice mechanisms of individual and institutions are poorly understood and are incorporated in a very macroscopic way as indicated by equation (15).

5. INTER-CITY TRANSPORT

The information presented in Fig. 1 indicated that inter-city transport expenditures in North America represented a relatively small proportion of the total expenditures on

transport. The bulk of inter-city passenger transport is by the private car. In a recent survey of inter-city trips (trips greater than 30 km) in Canada [37] it has found that about 85 percent of all person trips were by the private automobile. Of the 11 percent of the trips using public transport more than half were by air and about one-third were by bus. Rail passenger services accounted for only 1 percent of inter-city travel.

On the other hand the inter-city cargo movement costs represent about one-quarter of all transport expenditures with the bulk of these expenditures in the U.S.A. being on inter-city trucking. In Canada inter-city trucking has the largest share of the inter-city cargo market and the rail mode is relatively more important than in the U.S.A. capturing about one-third of the inter-city cargo transport expenditures.

In Canada the issues associated with inter-city transport have a distinctly different character to those associated with the urban transport sector. Most of the required inter-city transport infrastructure is in place and major capital investments in new capacity are not likely to be required for some time with the exception of some additional air terminal capacity at one or two locations along with some inadequacies in the Prairie grain export handling system. The inter-city transport sector problems are not concerned with shaping or controlling demand as is the case with the urban transport sector, but they are concerned, primarily with improving the utilization of existing transport services.

The urgency of improving the economic efficiency of inter-city transport in Canada is illustrated by Table 1 in which the annual economic costs of public infrastructure and the corresponding revenues for three transport modes are shown [37]. Rail costs are not shown in Table 1 since Canada's two major railways finance and operate their own infrastructure without direct government subsidy except in special cases. Table 1 shows that in 1975 the cost recovery varied from 17 percent in marine services to 59 percent for the highway transport mode.

A second major issue in Canada is the impact of transport services on development and the rail freight rates that should be charged for services to and from under-developed areas. The freight rate problem in Canada is complicated by many over-riding political questions that stem from the time that the transcontinental railways were constructed in the latter part of the 19th century. Economic development is very un-

Table 1. Annual Transport Infrastructure
Costs and Revenues for all Levels
of Government in Canada
(millions of 1975 constant dollars)

	Year		
	1968	1973	1975
AIR			
Annual Costs	340	436	516
Annual Revenues	95	137	152
% Cost Recovery	28%	31%	37%
MARINE			
Annual Costs	637	714	721
Annual Revenues	135	132	122
% Cost Recovery	21%	19%	17%
HIGHWAY			
Annual Costs	3330	4310	4796
Annual Revenues	2391	2696	2805
% Cost Recovery	72%	63%	59%

evenly distributed throughout Canada and many regions argue that their economic development has been compromised by the high transport costs associated with the long haul distance for goods between population centres.

5.1. Passenger Transport Analysis Tools

There has been very limited work on the development of inter-city travel demand forecasting techniques which are sensitive to changes in policy variables such as fares and levels of service. Some of the earliest work was conducted in connection with the Northeast Corridor Studies in the U.S.A. where extensions to the traditional gravity model were developed to estimate the impacts on modal demands of potential transport supply strategies [38]. Most planning studies, however, have used trend type forecasts of passenger travel demand because of the complexity of the factors influencing the inter-city travel demand market and the difficulties of forecasting future changes in these factors.

Recent passenger travel demand studies in Canada [39] have attempted to calibrate passenger travel demand models of the following type:

$$t_{iju}^m = t_{iju}(a_{iju}, c_{ij}^m) \cdot s_{iju}^m(c_{iju}^m) \quad (16)$$

where t_{iju}^m = The demand for travel by mode m between city pair $i-j$ at time u

$$t_{iju} = \sum_m t_{iju}^m$$

s_{iju}^m = The share of travel using mode m

a_{iju} = A vector of socio-economic activity variables

c_{iju}^m = A vector of attributes of transport mode m

One form of the modal share term of equation (16) that has been used is:

$$s_{ij}^m = e^{c_{ij}^m} / \sum_m e^{c_{ij}^m}$$

where the generalized cost of travel in the above equation,

c_{ij}^m , is similar to that described earlier in equation (14) for urban transport mode choice estimation.

5.2. Freight Transport Analysis Tools

Little effort has also been devoted to the development of analytical tools which might support freight planning studies. Most studies in Canada have used trend type forecasts and there has been little need for demand forecasting techniques which are sensitive to freight rate and level of service changes. In Canada Hutchinson et al [40] have used an adaptation of the urban travel demand forecasting process to develop an inter-city commodity flow model, while Hariton et al [41] have used time series analysis techniques to develop forecasts of major commodity movements by region.

One of the most elaborate attempts to develop an inter-city freight transportation forecasting model is that described by Bronzini et al [42]. With this approach an economic model uses transport price information to determine the origin-destination flows of commodities within a region. These flows are then split into flows by mode and a network simulation model assigns the mode-specific flows to network routes.

While these and other attempts to develop models of inter-city passenger and freight flows have provided some insight into the mechanisms underlying these demand patterns the techniques have not proved to be particularly relevant to current inter-city transport issues. It has been pointed out previously that in North America the principal concerns are with improving the economic efficiencies of existing modes rather than identifying new opportunities for investments in capacity. In Canada improved economic efficiencies are more likely to flow from changes in the regulatory environment rather than from additional technical analyses.

6. TRANSPORT AND REGIONAL DEVELOPMENT

It has been mentioned previously that one of the controversial transport-related issues in Canada is the extent to which transport services stimulate economic development in regions. While much has been written about this problem, particularly in the context of developing countries [43, 44] few suitable policy analysis tools exist. Some large scale models of economic development have been formulated and calibrated but they have found limited use in transport investment analysis. The more successful attempts have been

those concerned with the impact of transport on one specific sector such as agriculture [45].

Esguerra and Hutchinson [46] have described an approach to the identification of agricultural penetration road investments in Colombia. A linear programming formulation is used to model the probable responses of subsistence farms to the improved accessibilities to markets provided by new roads. As a farm's access to markets for cash crops is increased then it becomes worthwhile for a farmer to increase his production using existing production techniques and in some cases to introduce new production techniques. The model of the farm firm allows the value added by potential increases in agricultural production to be compared with the costs of road investments. A dynamic programming formulation is used to search for the optimal combination of agricultural penetration road investments.

A considerable body of literature exists on the impact of transport investments on development but it is beyond the scope of this paper to review this material. It is sufficient to note that much of the work is based very strongly in economic theory rather than the systems analysis tradition of the urban and inter-city transport sectors.

7. CONCLUDING REMARKS

Much of the work on transport systems analysis conducted during the past two decades has been concerned primarily with forecasting transport demands given estimates of future development patterns. The primary use of these estimates has been to assist in the identification of opportunities for new investments in transport capacity. Most of the modelling capabilities which have been developed in the urban and inter-city transport sectors consist of sequential sets of sub-models calibrated to cross-sectional data. While these types of models are capable of reproducing existing travel demands there are serious reservations about the capabilities of these models in forecasting marginal changes in transport demand over time.

In North America much of the urban and inter-urban transport infrastructure is in place and the planning problems are not concerned with handling future travel demands but are concerned with improving the economic efficiency of existing transport facilities and understanding the inter-relationships between transport and development. While some attempts have been made to model the transport sector in the much broader context of human activity distributions, available models are of limited value. Potentially the most productive area of research and development is improving our understanding of the

interactions between human activity distributions, transport demand and transport supply. Improved capabilities in this area will allow the spatial distributions of activities to be shaped so that the resulting transport demands will be more compatible with available transport services.

8. REFERENCES

- [1] Detroit Metropolitan Area Traffic Study, PART I - DATA SUMMARY AND INTERPRETATION, Detroit, Mich., U.S.A., 1955
- [2] Chicago Area Transportation Study, STUDY REPORT, Vols 1-3 Chicago, Illinois, U.S.A., 1959-1961
- [3] Metropolitan Toronto Planning Board, DRAFT OFFICIAL PLAN OF THE METROPOLITAN TORONTO PLANNING AREA, Toronto, Ontario, Canada, 1959
- [4] Thomson, J.M., GREAT CITIES AND THEIR TRAFFIC, Penguin Books, Harmondsworth, England, 1977
- [5] Hutchinson, B.G., PRINCIPLES OF URBAN TRANSPORT SYSTEMS PLANNING, McGraw-Hill Book Company, New York, U.S.A., 1974
- [6] Systems Analysis and Research Corporation, DEMAND FOR INTERCITY PASSENGER TRAVEL IN THE WASHINGTON-BOSTON CORRIDOR, Report NO PB 166884, National Technical Information Service, Springfield, Va., U.S.A., 1963
- [7] U.S. Department of Transportation, SUMMARY OF NATIONAL TRANSPORTATION STATISTICS, Washington, D.C., U.S.A., 1974
- [8] Canadian Transport Commission, TRANSPORT REVIEW: TRENDS AND SELECTED ISSUES, Research Branch, Ottawa, Canada, 1977
- [9] Voorhees, A.M., A GENERAL THEORY OF TRAFFIC MOVEMENT, Proceedings, Institute of Traffic Engineering, Vol 1, PP. 46-56, 1955
- [10] Wilson, A.G., A STATISTICAL THEORY OF SPATIAL DISTRIBUTION MODELS, Transportation Research, Vol 1, PP. 253-269, 1967
- [11] Hutchinson, B.G. and D.P. Smith, EMPIRICAL STUDIES OF THE JOURNEY TO WORK IN URBAN CANADA, paper submitted to Canadian Journal of Civil Engineering, 1978
- [12] Wilson, A.G., URBAN AND REGIONAL MODELS IN GEOGRAPHY AND PLANNING, John Wiley and Sons, London, England, 1974

- [13] Hutchinson, B.G. and D.P. Smith, MODELLING WORK TRIP DISTRIBUTION PATTERNS IN URBAN ONTARIO WITH CENSUS DATA, Research Report, Ministry of Transportation and Communications of Ontario, February, 1979
- [14] Blunden, W.R., THE LAND USE TRANSPORT SYSTEM, Pergamon Press, Oxford, England, 1971
- [15] Colston, M. and W.R. Blunden, ON THE DUALITY OF DESIRE LINE AND LAND USE MODELS, Proceedings, Australian Road Research Board, PP. 170-183, 1970
- [16] Black, J. and W.R. Blunden, MATHEMATICAL PROGRAMMING CONSTRAINTS FOR STRATEGIC LAND USE-TRANSPORT PLANNING, Proceedings, The International Symposium on Traffic and Transportation, Kyoto, Japan, 1977
- [17] Evans, S.P., A RELATIONSHIP BETWEEN THE GRAVITY MODEL FOR TRIP DISTRIBUTION AND THE TRANSPORTATION PROBLEM OF LINEAR PROGRAMMING. Transportation Research, Vol 7, PP. 39-61, 1973
- [18] Warner, S.L., STOCHASTIC CHOICE OF MODE IN URBAN TRAVEL: A STUDY OF BINARY CHOICE, Northwestern University Press, Evanston, Illinois, U.S.A., 1962
- [19] Stopher, P.R., A PROBABILITY MODEL OF TRAVEL MODE CHOICE FOR THE WORK JOURNEY, Research Record NO 283, Highway Research Board, 1969
- [20] de Donnea, F.X., THE DETERMINATION OF TRANSPORT MODE CHOICE IN DUTCH CITIES, Rotterdam University Press, Rotterdam, 1971
- [21] Charles River Associates, A DISAGGREGATED BEHAVIOURAL MODEL OF URBAN TRAVEL DEMAND, U.S. Department of Transportation, Washington, D.C., 1972
- [22] Domencich, T.A. and D. McFadden, URBAN TRAVEL DEMAND: A BEHAVIOURAL ANALYSIS, North Holland Publishing Company, Amsterdam, 1975
- [23] Stopher, P.R. and A. Meyburg, (eds.), BEHAVIORAL TRAVEL-DEMAND MODELS, Lexington Books, D.C. Heath and Company, Lexington, Mass.. U.S.A., 1976
- [24] Wardrop, J.G., SOME THEORETICAL ASPECTS OF ROAD TRAFFIC RESERACH, Proceedings, Institution of Civil Engineers, Part II, Vol 1, PP. 325-378, 1952

- [25] Van Vliet, D., ROAD ASSIGNMENT - I : PRINCIPLES AND PARAMETERS OF MODEL FORMULATION, Transportation Research, Vol 10, PP. 137-143, 1975
- [26] Van Vliet, D., ROAD ASSIGNMENT - III : COMPARATIVE TESTS OF STOCHASTIC METHODS, Transportation Research, Vol 10, PP. 151-157, 1976
- [27] Nguyen, S., UNE APPROCHE UNIFIÉE DES MÉTHODES D'ÉQUILIBRE POUR L'AFFECTATION DU TRAFFIC, Publication NO 171, Département d'Informatique, Université de Montréal, 1974
- [28] Florian W. and S. Nguyen, AN APPLICATION AND VALIDATION OF EQUILIBRIUM TRIP ASSIGNMENT METHODS, Transportation Science, Vol 10, NO 4, November, 1976
- [29] Lowry, I.S., A MODEL OF METROPOLIS, RM-4035-RC, Rand Corporation, Santa Monica, Calif., U.S.A., 1964
- [30] Batty, M., URBAN MODELLING: ALGORITHMS, CALIBRATIONS, PREDICTIONS, Cambridge University Press, Cambridge, England, 1976
- [31] Hutchinson, B.G., LAND USE-TRANSPORT MODELS IN REGIONAL DEVELOPMENT PLANNING, Socio-Economic Planning Sciences, Vol 10, PP. 47-55, 1976
- [32] Sarna, A.C. and B.G. Hutchinson, A DISAGGREGATED LAND USE-TRANSPORT MODEL FOR DELHI, INDIA, Transportation, 1979
- [33] Said, G. and B.G. Hutchinson, AN URBAN SYSTEMS MODEL FOR THE TORONTO REGION, 2nd International Conference on Large Scale Engineering Systems, 1978
- [34] Sharpe, R., J.F. Brotchie, P.A. Aherne and J.W. Dickey, EVALUATION OF ALTERNATIVE GROWTH PATTERNS IN URBAN SYSTEMS, Computing and Operations Research, Vol 1, PP. 354-362, 1974
- [35] Hopkins, L.D., LAND USE PLAN DESIGN : QUADRATIC ASSIGNMENT AND CENTRAL FACILITY MODELS, Environment and Planning A, Vol 9, PP. 625-642, 1977
- [36] Gupta, J.D. and B.G. Hutchinson, REGIONAL DEVELOPMENT OPTIMIZATION FOR THE DELHI REGION OF INDIA, Environment and Planning A, 1979

- [37] Transport Canada, TRANSPORTATION : A NATIONAL AND REGIONAL PERSPECTIVE, Discussion Paper, First Ministers Conference on the Economy, November, 1978
- [38] Quandt, R.E. and W.J. Baumol, THE DEMAND FOR ABSTRACT TRANSPORT MODES : THEORY AND MEASUREMENT, Journal of Regional Science, Vol 6, 1966
- [39] Rea, J.G., M.J. Wills and J.B. Platts, CANPASS : A STRATEGIC PLANNING CAPABILITY FOR INTERCITY PASSENGER TRANSPORTATION, Strategic Planning Group, Transport Canada, March, 1977
- [40] Hutchinson, B.G., W.B. O'Brien and I.N. Dawson, A NATIONAL COMMODITY FLOW MODEL, Canadian Journal of Civil Engineering, Vol 2, NO 3, PP. 292-304, 1975
- [41] Hariton, G., R. Lee and U. Zohar, ECONOMETRIC FORECASTING MODEL : DEMAND FOR FREIGHT TRANSPORT IN CANADA, Research Branch, Canadian Transport Commission, Ottawa, Canada, 1976.
- [42] Bronzini, M.S., J.H. Herendeen, J.H. Miller and H.L. Womer, A TRANSPORTATION SENSITIVE MODEL OF A REGIONAL ECONOMY, Transportation Research, Vol 8, PP. 45-62, 1974
- [43] Fromm, G., (ed.), TRANSPORT INVESTMENT AND ECONOMIC DEVELOPMENT, Transport Research Program, The Brookings Institution, Washington, D.C., 1965
- [44] Wilson, G.W., et al, THE IMPACT OF HIGHWAY INVESTMENT ON DEVELOPMENT, Transport Research Program, The Brookings Institution, Washington, D.C., 1966
- [45] International Bank for Reconstruction and Development, SYSTEMS ANALYSIS OF RURAL TRANSPORTATION, Economics Department Working Paper, NO 77, 1970
- [46] Esguerra, G. and B.G. Hutchinson, EVALUATION OF DEVELOPMENT TYPE TRANSPORT INVESTMENTS, 8th Annual Congress, The Engineering Institute of Canada, 1974

THE PRACTICE OF MILITARY OPERATIONS RESEARCH

DAVID A. SCHRADY

Naval Postgraduate School
Monterey, Ca. 93940, U.S.A.

ABSTRACT. After presenting a brief discussion of its origins and development, the paper is concerned with the practice of military operations research in terms of its proper role in defense decision making, the people who do analysis, the difference between analysis and advocacy, and the clients of military operations research.

1. INTRODUCTION

Operations research is what many of us teach, do, or use daily in our professional lives. Indeed, I should say "lives" without qualifying the employment of operations research to professional matters exclusively. Some may argue that operations research affects them only in their professional dealings and that common sense governs their personal and family lives. Operations research is, after all, a scientific discipline involving mathematics, probability and statistics, stochastic processes, optimization techniques, computing and simulation, economics and natural laws. Still some describe operations research as simply "quantified common sense."

I am grateful for the opportunity to address this conference, the first international conference on operations research ever to be held in Korea. The title of my remarks, "The Practice of Military Operations Research," suggests that I am more concerned with how operations research is done and how it affects defense decision making than with specific applications or techniques. I fundamentally believe that operations research as a discipline is a form of common sense, a common sense that has been affected by mental discipline in the scientific method in general, and in economics, the laws of probability, and optimization techniques in specific.

I once advised a U.S. Naval officer who was at best a marginal student. At the midpoint of his education at the Naval Postgraduate School (NPS), he was terribly unsure whether he should go on what we call an "experience tour" or stay at NPS for the six weeks and try to remedy his academic deficiencies. The experience tour is a time when students go out to an organization doing operations research and participate as analysts to gain some "real world" experience. We waited until his examination scores were ready and, at the eleventh hour, I talked him into going on the experience tour.

The officer enjoyed the experience tour and produced, in the six weeks allotted, a finished piece of analysis complete with recommendations for policy changes. The changes this officer recommended were adopted at his experience tour site and that organization recommended that these changes be adopted command-wide. This officer's experience tour was by

any measure very successful. While the officer gained a new appreciation of his own talents and capabilities, he stated that his analysis and recommendations involved only common sense. He went on to say that it was a little disturbing because his common sense was no longer what it had been. His formal education in operations research had altered forever the fundamental nature of his common sense. I view this contribution of operations research as perhaps of fundamental importance in the practice of military operations research. However, before going further with the notion of practice, it is instructive to survey the origins and development of the field.

2. BACKGROUND AND HISTORY

Wartime, as well as politics, can create strange bedfellows. In the case of operations research, World War II provided the impetus for a fundamental change in the relationship between scientists and the affairs of man. A self-respecting physicist, especially if he was educated in the 1930's, knew what activities were worthy of a physicist and those that weren't. Nuclear physics was respectable. Solid state physics was at the outer fringe of respectability. Management consulting sorts of activities were definitely not respectable. When wartime conditions dictated enlisting scientists -- physicists, chemists, and mathematicians -- to help out with operational military problems, it was a significant historical event.

Some may note that Archimedes' advice was sought by Hieron, King of Syracuse in the matter of breaking naval sieges during the Second Punic War. The well-read may even recall Benjamin Franklin's light-hearted letter in 1775 to Joseph Priestly, anticipating by 200 years his country's problems in quantifying American experience against the Viet Cong. Franklin wrote: "Britain, at the expense of three million, has killed 150 yankees this campaign which is ~~X~~ 20,000 a head and at Bunker Hill she gained a mile of ground, half of which she lost by our taking post on Ploughed Hill. During the same time 60,000 children have been born in America. From these data any mathematical head will easily calculate the time and expense necessary to kill us all, and conquer our whole territory." Finally, it must be noted that during World War I, Viscount Tiverton did

significant work that would now be called operations research for the Royal Naval Air Service, and that Fredrick W. Lanchester published his work on the relationships between victory, numerical superiority, and fire power.

Despite the earlier work, World War II signaled the emergence of operations research and the context was military. Early operations research was characterized as the employment of scientists in the analysis of operations. Analysis was focused upon the present, that is, current operations and the effective use of current systems. What was significant in the employment of scientists? A priori, scientists were individuals with considerable intellect and a strong training in the scientific method. What is so wonderful about the scientific method you may ask?

One can neatly dichotomize problem solving into that which comes from either scientific inquiry or unscientific inquiry. Unscientific inquiry is sometimes referred to as common sense. Scientific inquiry is characterized by rationality, empiricism, and rigor. These are powerful tools and scientific inquiry was very successful when applied to operational military problems. Those who may doubt the overwhelming superiority of scientific inquiry should read some of the literature written since the expiration of the thirty year moratorium imposed by the British Official Secrets Act. R. V. Jones [1] documents clearly the futility of investigative committees chaired by politicians, the destruction of purposeful investigation wrought by ego and personal gratification, and the nearly superstitious beliefs of some leaders of the armed forces. Examples of the latter were the refusal of RAF Bomber Command to order crews to stop trying to use their IFF equipment to jam enemy radars while over enemy territory and a nearly two year delay in the employment of chaff as a bomber penetration aid.

In summary, operations research was born out of wartime necessity. It employed persons from the sciences, both for the scientific method they brought to operational problem solving and because World War II saw the early employment of radar, radio direction finding and other manifestations of technology for which scientific understanding was required. Finally, the wartime operations analysts focused their attention on real problems involving current operations and the employment of current systems. Further wartime analyses were generally performed direct for those who controlled the operations and therefore could implement the recommendation.

3. DEVELOPMENT

On the U.S. side, the post-war period saw the U.S. Navy continue in the operations research business. The establishment of the U.S. Air Force saw most of the Army analysts depart from that Service and go with what had been the Army Air Corps. The U.S. Army was slow to institutionalize operations research but (getting a little ahead of my story here) Mr. McNamara made it a matter of survival for them in the mid-1960's. Many of those who were uniformed analysts during the war went back to their civilian careers in business, industry, and academia, and they took with them the conviction that operations research would prove as useful in business and industry as it had been in defense. In 1951, the Naval Postgraduate School and Columbia University began the first graduate academic programs in operations research, the program of the Naval Postgraduate School having a strong military orientation. During the 1950's, operations research began to have an impact in the private sector in inventory control, production planning, marketing, scheduling, problems of congestion, and optimal employment of resources.

Early operations research was frequently interdisciplinary. While the virtues of an interdisciplinary approach are often praised, in World War II, operations research was interdisciplinary by necessity. Once the academic world started educating operations research students, the situation changed somewhat. As an aside, Harlan Cleveland, formerly President of the University of Hawaii, notes that interdisciplinary courses are normally "team taught." That is, each faculty member lectures on his discipline and it is the student that is expected to be interdisciplinary!

Still, the interdisciplinary notion leads us to a fundamental change which occurred in military operations research in the early 1960's. That change is the emergence of the economist as a contributor to or even dominant participant in military operations research -- and the invention of the term systems analysis. Economics had always played a role in operations research. The full title of the John VonNeumann and Oskar Morgenstern classic study of game theory was "The Theory of Games and Economic Behavior" and the premier inventory control book of the 1950's, "Studies in the Mathematical Theory of Inventory and Production" was coauthored by the Nobel Prize economist Kenneth Arrow.

Indeed, one of the old alternative definitions of operations research was "economics rediscovered by physicists."

There is no doubt that the economist has become a key participant in military operations research and systems analysis at the national level. Dr. B. O. Koopman, an original U.S. Navy Operations Evaluation Group member and developer of the theory of search, complains bitterly that the economists came to operations research not to join the team but to take charge of it. Koopman also describes the economist as one who knows the cost of everything and the value of nothing.

There is one other change that has occurred in military operations research at the national level. The focus is planning, the setting of requirements, and the evaluation of possible future systems and operations. In its focus on planning, military operations research is frequently done for persons or agencies that do not control the operations or even the planning. This analysis faces a difficult or impossible set of hurdles to its implementation. In the 1960's when systems analysis was the medium in the Department of Defense, the Armed Services rushed their analytical assets into the planning and program arena in Washington. Today the Armed Services have more analytical assets and there is a better balance between planning and programming activities in Washington and the tactical analysis activities in the field.

Today's focus of military operations research includes logistics, manpower (which accounts for over 50% of the U.S. military budget), doctrine, force structuring, tactical analysis, test and evaluation, intelligence, and strategic planning. I note that analysis of weapons systems and tactics is the subject of the majority of the papers to be presented subsequently in this session. For all its limitations, operations research/systems analysis has an enormous impact on defense decision making. It is not a matter of liking or not liking analytical methods as a basis for defense decision making. It is simply superior to other forms of inquiry.

Having set the stage perhaps at too great a length, it is time to discuss the practice of military operations research in terms of its proper role in defense decision making, the people who do analysis, the difference between analysis and advocacy, and finally the clients of military operations research.

4. THE ROLE OF ANALYSIS

The role of operations research in decision making is to assist, as best it can, those who must decide by providing an adequate description of the problem and by structuring and evaluating alternative solutions. Tom Saaty suggested that "operations research is the art of giving bad answers to problems to which otherwise worse answers are given." Saaty's glib definition is both pessimistic and truthful, because operations research is rarely able to provide the correct solution to any problem. As with science in general, the power of operations research is achieved through quantification. The adequacy of any analysis is strongly correlated with the extent to which all aspects of the problem lend themselves to measurement and quantification. The rub is, of course, that almost all real problems have aspects which, if considered at all by the analyst, can only be treated qualitatively. Thus analysis is properly only a part of the material from which final decisions must be made. To believe otherwise is to be arrogant and foolish.

It is said that a major figure in the U.S. Defense Department in the 1960's, a figure whose background was devoid of military experience, initially rejected the opinions and inputs of senior military officers, believing that all issues of choice were susceptible to complete quantification. Seven years later this official, still in the same important position, admitted that his earlier perceptions were not totally correct and that military judgement was relevant. As a postscript to this admission, the official added that while he now realized the necessity of incorporating military judgement the only military judgement he would accept as input was his own.

In clear recognition of the limitations of analysis, the preface to the 1979 CNO Studies and Analysis Program states: "This study program addresses the most significant issues for which analysis offers reasonable expectation of providing valuable assistance in improving Navy program planning." In summary, operations research can be most valuable in decision making when both the analyst and the decision maker understand that analysis is only a part of the material from which a decision must be drawn. Depending upon the problem, analysis can be the most important ingredient or it can play only a minor role. Not all examples of bad analysis are a

criticism of the analyst. In many cases, the criticism should be laid to a naive client.

5. THE ANALYSTS

The first operations analysts were civilian scientists who were drafted or otherwise associated with the armed services in a period of grave national crisis. When World War II ended they returned to their civilian careers. The U.S. Services thus had decisions to make as to in-house analytical capabilities and the extent to which their in-house capabilities would include uniformed analysts. At the present time, all U.S. Services have in-house analysis shops at the headquarters level and in the field. All of the Services have officer personnel management systems that recognize a primary (warfare) specialty and a secondary specialty in operations research/systems analysis. Presently the Army required approximately 600 officers to have an operations research/systems analysis secondary specialty, the Navy about 400, the Marine Corps about 150, and the Air Force 200+. Additionally the Services, the Office of the Secretary of Defense, the Office of Management and Budget, the General Accounting Office, and the Congress all employ civilian analysts and all are participants in defense decision making.

An assertion made earlier in this paper concerned the superiority of scientific inquiry. It follows then that those persons doing analysis should possess an education in the sciences, broadly defined, and should probably have graduate education. There are virtually no undergraduate operations research education programs in the U.S. The Center for Naval Analyses has a professional staff, 84% of whom have a graduate degree, consisting of operations researchers, physicists, chemists, mathematicians, engineers, economists, political scientists, and sociologists.

Still the term "analyst" is much abused. While not being an infalible rule, it seems prudent to require that a person claiming to be an analyst possess a graduate education in operations research/systems analysis or one of the useful sciences. It also seems crucial, and the Services agree, that a fair proportion of the analysts available to a Service be uniformed members of that Service. It is the career officer who has the operational knowledge so crucial

to adequate analysis for defense decision making; having process knowledge and analytical capabilities in the same individual is a powerful arrangement. The uniformed analyst is also crucial in judging the adequacy of analyses done by outsiders for his Service. Finally the uniformed analyst will become, with time and seniority, the client who will task others to do analyses. I am proud to note that over the past twenty-eight years, the Naval Postgraduate School has educated uniformed analysts for the U.S. Navy, Army, Air Force, Marine Corps, Coast Guard, and officers from over two dozen allied nations, including nearly two dozen officers of the Republic of Korea Armed Forces.

6. ANALYSIS AND THE ADVERSARY PROCESS

One likes to think of the scientific method as being rational, open, explicit, rigorous, and verifiable. One hopes further that as a scientist, the analysts will be impartial, honest, and truthful. However, recall the prophecy of Morse and Kimball [2] in stressing that analysis be done only for clients who are in command and who are capable of making decisions. They further noted that; "The analyst must never denigrate into a salesman for a laboratory or a service branch." The Operations Research Society of America [3] noted, "It may happen that studies are sponsored by an authority who cannot make the final decisions under consideration. In this case, the analyses may be used not only to arrive at, but also to justify the authority's recommendation. Then the operations analyst may be placed in an advocacy position."

The adversary process is a decision making process widely accepted in our society, but it operates by rules that are somewhat different from those of the scientific method. The adversary process requires that participants attempt to influence decisions through presentation of their cases. The adversary process requires that what the participant asserts should be true, but it does not require that it be the whole truth. The competent analyst must clearly realize the difference between analysis and advocacy. I do not mean to champion analysis and condemn advocacy. Both are useful but a good analyst must recognize the difference.

The 1969 debate before the Congress on ballistic missile defense, the ABM debate, focused the issues of analysis and

advocacy. Senator Henry M. Jackson, in the report of the Subcommittee on National Security and International Operations commented about the analyses brought to the Committee's attention:

"Analysis, of course, varies greatly in quality. One often wishes that advisors with different points of view would confront each other directly and in public so that hidden or unstated assumptions could be revealed and the different modes of analysis explored."

Because of the potential harm that the situation described by Senator Jackson could do to the profession, the Operations Research Society of America published a report [3] entitled "Guidelines for the Practice of Operations Research." Analysts, as well as their clients, should be familiar with this report. The report contains the following excerpt:

"The analyst, as an analyst, must restrict his analysis to the quantifiable and logically structured aspects of the problem only. In complex problems, perhaps the most valuable thing the analyst can do is to point out to his client that there are uncertainties inherent in his analysis and their conclusions, uncertainties deriving from such factors as:

- Lack of agreement on means of evaluating the worth of complex systems;
- Uncertainty about the technical capabilities and costs of systems yet unbuilt;
- Uncertainties about environmental and operational factors that influence performance;
- Uncertainties about the future capabilities and intentions of possible adversaries."

During World War II there were sometimes opportunities for the analyst to demonstrate the confidence he had in his analysis. The story is told about the analyst who recommended that bombers flying against the Japanese fly at 9-10 feet, in the "seam" between their medium and high altitude antiaircraft guns. In that case, it was possible for the analyst to volunteer to fly in the lead bomber. In the planning of future systems and future force structures, few such opportunities for the analyst to demonstrate his honesty, unbiasedness, and objectivity exist.

7. THE CLIENTS OF MILITARY OPERATIONS RESEARCH

This discussion of the practice of military operations research would not be complete without a discussion of the qualities and responsibilities required of the clients of such analyses. The client could be a program, a warfare specialty, a Service, the Department of Defense, the Congress, or the President. A good client is a sophisticated one who knows the power and the limitations of analysis. A good client is one who does not provide the answer and ask the analyst to prove it. A good client asks operations research to help with only those problems which are reasonably amenable to structuring and quantification. It is wrong for a client to request analysts to determine an unqualified, cardinal measure of total defense capability. It is wrong to ask by how much that capability will increase if system X is added to the arsenal. It is wrong to ask operations researchers to predict the future. Good analysis results from good work by good analysts as a result of good tasking from competent clients.

8. SUMMARY

The successful practice of military operations research depends upon a number of factors other than models, algorithms, and efficient computer codes. People are important as are the organizational arrangements. Analysts are persons who at a minimum have graduate education in a scientific discipline and who understand the operations they analyze. Ideally, studies and analyses are performed for the persons or agencies who have the authority to make decisions. While this is clearly not totally feasible in the context of defense planning at the national level, it should be the rule in the analysis of operations. All parties, analysts and clients alike, should appreciate the role of operations research as an aid to decision making and not a substitute for executive decision making. The role of analysis is to provide a better understanding of the process or problem and to provide the decision maker with evaluations of alternative solutions. Finally, the profession and the military are best served when the analyst functions as an analyst and avoids advocacy

9. REFERENCES

- [1] Jones, R.V., THE WIZARD WAR, Coward, McCann, and Geoghegan, Inc., 1978.
- [2] Morse, P.M. and G. E. Kimball, METHODS OF OPERATIONS RESEARCH, The MIT Press, 1951.
- [3] "Guidelines for the Practice of Operations Research." OPERATIONS RESEARCH, Volume 10, Number 5, September 1971.

DRAFT

RESOURCE ALLOCATION AND DEFENSE PLANNING
IN RETROSPECT AND PROSPECT*

CHARLES WOLF, JR.

The Rand Corporation
1700 Main Street
Santa Monica, California 90406, USA
February 12, 1979

ABSTRACT. A retrospective view of resource allocation and defense planning is presented as a review of a landmark book on the subject, written nearly twenty years ago: *The Economics of Defense in the Nuclear Age*, by Charles Hitch and Roland McKean, 1960. Although Hitch-McKean have met remarkably well the test of time, a number of omissions and commissions are discussed from the vantage point of the present.

In examining prospective issues of resource allocation and defense planning, several topics are addressed: increasing competition for public sector resources; possibly increasing real costs in defense industry in the midst of an inflation-prone economy; the growing relevance of energy policy in defense planning; the new importance of foreign exchange markets and exchange rate uncertainty in the planning and deployment of forward based forces; and the new opportunities provided by technological developments for capital-labor substitutions in the planning of defense forces.

Finally, the paper suggests that implementation analysis will and should receive greater attention in future defense planning studies.

* An invited talk prepared for the Pacific Conference on Operations Research, April 23-28, 1979, Seoul, Korea.

1. INTRODUCTION

The topic which the organizers of this conference asked me to address is formidable in scope. The first part of the topic, "Resource Allocation," embraces literally all of defense economics. The second part, "Defense Planning," if interpreted literally, covers strategic forces, doctrines, and targeting; general purpose forces and their employment; defense research, development and systems acquisition; command, control and communications, the effects of SALT II and other arms control agreements on all of the foregoing; and so forth.

Furthermore, the combination of "resource allocation" and "defense planning" implicitly covers other special policy issues, as well. For example, such issues as arms transfers, and the structure, scale, and role of overseas bases and deployments of U.S. forces, in Korea and the Philippines, also come within the topic, because they involve the allocation of defense resources. Even an issue as remote as U.S. export control policies legitimately comes within the purview of resource allocation and defense planning. For example, U.S. exports of computer technology may affect Soviet capabilities, and hence influence U.S. defense planning and resource allocations.

Frankly, I don't know anyone who is qualified to address all of these issues adequately without drawing upon a substantial number of coauthors! Certainly, I don't feel qualified to do so.

What, then, should one do in facing such a formidable subject, under severe limitations of time and knowledge and with a reluctance to draw upon a dozen or more colleagues at Rand or elsewhere to be coauthors?

My response is to attempt only a very limited treatment of the topic.

The retrospective part of my remarks will review a landmark book in this field, *The Economics of Defense in the Nuclear Age*, written by my former Rand colleagues, Charles Hitch and Roland McKean, and published in 1960.¹ This book is probably the most comprehensive work on resource allocation and defense planning published in the past two decades. The question I will address in my retrospective remarks is:

1. Charles Hitch and Roland McKean, *The Economics of Defense in the Nuclear Age*, Harvard University Press, Cambridge, Massachusetts, 1960.

How does the Hitch-McKean work look nearly twenty years after publication? What are its significant omissions and commissions?

The prospective part of my remarks will then try to identify and comment on a few major current and impending issues of resource allocation and defense planning.

Before I proceed, let me apologize for concentrating on American defense economics and planning issues. I have not attempted a broader treatment from the viewpoint of the other countries and regions represented at this conference, although such a broader treatment would be desirable. I hope you will be able to link the observations I will make to defense issues in other countries with which you are familiar.

2. IN RETROSPECT

Let me begin by refreshing your recollection of the Hitch-McKean book. It is divided into three sections: Part I. "Resources Available for Defense," which addresses the relationship between defense and the domestic economy; Part II. "Efficiency in Using Defense Resources," dealing with allocative options within the military sector, and the competing and complementary relations among operations decisions, procurement and force composition decisions, and research and development decisions; and Part III. "Special Problems and Applications," which covers such specialized topics as the economics of military alliances, logistics, economic warfare, and R&D decisionmaking.

In my judgment, the Hitch-McKean book (hereafter referred to as H-M) stands up remarkably well to rereading in 1979. Nevertheless, there are a number of interesting omissions and commissions and differences in emphasis that, in retrospect, appear worthy of comment.

1. In discussing the relationship between inflation and defense, H-M stress the effect of defense spending on inflation. The authors don't give much attention to the effect of inflation on defense spending. From the vantage point of 1979, the latter relationship is at least as important as the former. The deep-seated and persistent inflationary characteristics in the American economy at present have major consequences for defense resource allocation and planning. For example, the differential rate at which manpower costs and capital costs have been rising should affect opportunities for efficient substitutions between capital and labor in planning, developing, and operating weapons systems and military forces.

Consider, also, the political commitment made by the U.S. to our NATO allies to raise the real value of defense spending on NATO-related forces. The issue of a proper deflator for defense expenditures, to calculate real as against nominal outlays, has now become an important issue of defense planning and policy. It was largely irrelevant to defense planning in the relatively stable economic environment in which H-M was written.

2. The index to H-M has no reference to "energy" or "oil." However, the chapter dealing with "economic warfare" includes this prescient reference:

"Control over Middle Eastern oil by any single power or bloc would be a comparatively potent weapon because it could upset the Middle East and European economies sharply, particularly during an initial period before adjustments could be made."²

Notably, the quotation omits the U.S. However, with U.S. oil imports currently amounting to almost 50 percent of national consumption, the U.S. has also become highly vulnerable to a protracted interruption of Middle Eastern oil supply, although less vulnerable than our European and Asian allies. One major consequence of this sharp economic change is the importance for national security of the proposed Strategic Petroleum Reserve. This \$25 billion oil stockpile has become a major issue and resource claimant in U.S. defense and foreign policy planning in the 1980s, as it was not in the 1960s.

3. The index to H-M contains no reference to such major issues of international finance as those related to the roughly \$500 billion of Eurodollar overhang in international exchange markets. One result of the overhang is a high degree of instability in foreign exchange rates under the present flexible rate system. The fluctuating value of the dollar creates a serious problem of resource allocation for U.S. forward-based forces in Europe, Korea, and the Western Pacific. These were not important issues at the time H-M was written. They are serious concerns now and likely to become more so in the years ahead.

2. Op. cit., p. 303.

4. H-M do not address the subject of nuclear proliferation, perhaps one of its more surprising omissions.

Expansion of nuclear reactors in the past two decades--probably a considerably greater expansion than strictly economic calculations would have warranted--has created serious issues for defense planning in the 1980s that were not foreseen in the 1960s. On the one hand, these issues relate to the entire fuel cycle: uranium supply; enrichment and re-processing technology and facilities; waste management; and the breeder reactor. How these stages of the cycle are managed or avoided will affect the supply of weapons-grade nuclear materials in various countries and regions. On the other hand, proliferation also depends on the incentives that exist, or may be perceived to exist, for such countries as Korea and Taiwan to acquire nuclear weapons in order to strengthen deterrent capabilities they may feel have been weakened by changes in U.S. force posture or foreign policy.

5. At a more microeconomic level, one finds surprising the absence in H-M of any detailed discussion of policies and costs relating to military manpower. Since establishment of a volunteer military force in 1973, a close linkage has been created between civil labor markets and the market for military manpower. The resulting impact on the budgetary costs of defense, and the optimal structuring and operation of forces as between capital-intensive and labor-intensive components, have become important issues of defense planning in the 1980s, which were not of concern in the 1960s.

6. In some respects, H-M reflect a view of the nature of war, of deterrence, and of defense economics that is less complete and sophisticated than most of us have, or think we have, currently. For example, H-M make the following remarkable observation:

"In our view the problem of combining limited quantities of missiles, crews, bases, and maintenance facilities to 'produce' a strategic air force that will maximize deterrence of enemy attack is just as much a problem of economics (although in some respects a harder one) as the problem of combining limited quantities of coke, iron, or scrap, blast furnaces, and mill facilities to produce steel in such a way as to maximize profits."³

3. Op. cit., p. 2.

From the vantage point of 1979, most of us would feel that this statement ignores several considerations which seriously degrade the analogy: the vastly greater difficulty of specifying the "deterrence" objective than the "profit" objective; the fact that "deterrence" depends on the perceptions and the actions of Soviet decisionmakers as well as on U.S. actions; and the potentially important interactions between U.S. decisions about developing, procuring and deploying forces, and those of the Soviet Union.

Notwithstanding these few critical comments, I repeat my earlier general assessment: *The Economics of Defense in the Nuclear Age* remains, twenty years after it was written, a valuable survey of resource allocation and defense planning. I think any of us would be happy if work we have done passes the test of time as well as does H-M!

H-M make a particular point which is unusually prescient with respect to the economic issues related to defense planning in the future. Written at a time when nuclear weapons were often accorded an exaggerated, and sometimes exclusive, role in defense analysis, H-M remind the reader of the importance of "economic strength as a deterrent of lesser aggression."⁴ The importance they ascribe to mobilization potential--the ability to boost defense spending, and undertake rapid economic and military mobilization efforts--as a deterrent to conflicts short of all-out nuclear attack, is unusually discerning. I believe this issue of mobilization potential will acquire increased interest among the matters that defense planners are concerned with in the years ahead.

In recalling the importance of mobilization capabilities in the past, I think H-M envisaged the future.

3. IN PROSPECT

I will now address a few current issues of defense economics and defense planning in the U.S. that are likely to assume greater importance in the years ahead.

3.1. Defense Resource Allocation and the U.S. Economy

At the macroeconomic level, a number of important changes are under way in the relationships between defense resource allocation and the economy as a whole.

4. Op. cit., p. 317.

For example, over the past twenty years the proportion of U.S. resources devoted to defense has declined sharply. Defense expenditures declined from 12 percent of national income in 1957 to 11 percent in 1967, 6 percent in 1977, and less than 5 percent in the proposed 1980 budget. Defense planning will face tighter resource constraints than in the past. Consequently, more careful analysis of alternatives, tradeoffs, and costs and effectiveness of competing ways of allocating defense resources, will be needed in the future.

In the U.S., but perhaps not in Korea and Japan, we have entered a period in which American taxpayers and their representatives are displaying a growing resistance to expenditures by the public sector. In part this resistance arises from a general disenchantment with programs undertaken by government, perhaps even more in non-defense than in defense sectors. The result is additional constraints on resource availability in the public sector.

How will these fiscal limitations affect resource availabilities for defense purposes? The question is both important and difficult to answer. In the President's proposed budget for FY 1980, defense is the only major federal government program for which a growth in real expenditure is planned. Expenditures on social programs, energy, transportation and housing, are, generally, scheduled for constant or reduced real outlays.

Will this small expansion in defense outlays be enduring or transitory?

There are a number of reasons why it may be transitory: for example, the short-run political maneuvering which perhaps links increased defense spending to lining up Congressional support for a SALT II agreement; the temporary pledge to our NATO allies to raise NATO-related expenditures by the U.S. to match the planned increases by other NATO members; and the political pressures in some circles to curtail defense outlays in order to expand certain social programs.

On the other hand, there are reasons why the increases in defense spending may persist. For example, the increased resources allocated to the defense sector are, in part, a consequence of the Soviet build-up in both strategic and general purpose forces, and there is as yet no evidence that these will let up. Moreover, it can be argued that the American taxpayer's resistance to public expenditures, as reflected in Proposition 13 and other similar measures on the American legislative scene, may be directed more toward non-defense than defense programs, because non-defense programs undertaken by the public sector often entail activities in which the private sector might plausibly assume a

greater role (such as in housing, transportation, and energy) if public sector programs were reduced. This argument does not apply in the defense sector, the classic case of a "pure" public good which the market for "private" goods and services cannot replace.

Changes that have occurred in the structure of American industry may lead to reduced price competition in U.S. industry, including defense industry. These changes result from the greater market power of both labor and business, leading to stronger cost-push inflationary forces, which are reinforced by the cost impact of regulatory and environmental constraints. Defense costs may therefore rise at the same time as defense budgetary appropriations are under severe constraints.

On the other hand, an optimist might argue that a leaner and more efficient defense industrial base may result. Tighter budgets and rising costs may create pressures toward economizing, streamlining, and increased efficiency among the surviving firms.

This scenario is certainly possible, but I think unlikely.

More likely, over the next few years, is a scenario in which the rate of innovation and productivity increase in American industry continues to be low. In recent years, productivity increases have fallen from approximately three percent per year to about one percent. The result may be a lower price elasticity of supply of defense resources in the future than in the past. Resource mobilization in the American economy may become increasingly, perhaps excessively, costly. Hence, mobilization may become politically less feasible in the future.

Defense planners, to the extent that they are concerned with planning for possible "surge" expansion of the defense sector, will have to take these new structural developments prominently into account.

3.2. The Defense Sector and the International Economy

I will address only two aspects of the changing relationships between the defense sector and the international economy.

As noted earlier, the U.S. is now more vulnerable to an oil embargo, or the threat of embargo, than in the past. It is a deplorable commentary on U.S. policy making that the prospect for reducing this dependence over the next three to five years looks dim.

Two important implications follow for defense planning. The strategic petroleum reserve should be viewed as an

important aspect of defense planning and defense policy. Whether the reserve covers import demands for 120 days, as was originally intended, or only 60 days, will seriously affect resource mobilization problems in the event of an emergency.⁵ And energy policy, in general, will be an important aspect of defense planning and defense policy in the future.

Future defense planning will also be affected by the huge amounts of dollars that are held abroad. At the present time, foreign dollar holdings exceed the total U.S. money supply by about 20 percent! Small changes in the confidence and expectations of these asset holders can have dramatic effects on the exchange value of the dollar, and hence on the costs of forward deployed military forces. I expect that this source of enhanced uncertainty will become increasingly important to defense planners.

3.3. Technological Development in Defense Planning

Recent and impending developments in information processing, guidance, and sensor technology will have dramatic implications for defense planning and resource allocation. On the one hand, the new technology makes possible more complete and accurate command and control of the battlefield, as well as more accurate targeting and delivery of ordnance. On the other hand, the rising budgetary costs of manpower, resulting from the all-volunteer force and the resulting link between military compensation and the civil sector labor market, creates a greater incentive for defense planners to save on labor costs in force posture and system development decisions. As a result, defense analyses in the future will have to give more explicit attention to capital-labor substitutions in the development of systems, and in the structuring and operation of forces.

4. CONCLUSIONS

The foregoing list is not exhaustive, but it indicates some of the major issues affecting resource allocation in defense planning that lie ahead: increasing competition for public sector resources; possibly increasing real costs in defense industry in the midst of an inflation-prone economy;

5. In fact, the existing petroleum reserve is below the lower of these two levels.

the growing relevance of energy policy in defense planning; the new importance of foreign exchange markets and exchange rate uncertainty in the planning and deployment of forward based forces; and the new opportunities provided by technological developments for capital-labor substitutions in the planning of defense forces.

I will conclude with one observation relating to the methodology of defense policy and planning studies.

In defense planning studies in the future, we will have to give greater attention to implementation analysis than we have in the past. Typically, planning studies have proceeded by comparing the costs and effectiveness of alternative programs, employing a more or less formal model of the problem under consideration. A preferred program is then selected by applying the usual sort of criterion to the results of the model: for example, maximizing effectiveness for a specified budget, or minimizing costs for specified effectiveness. Sometimes, indeed more and more frequently, a dominant choice doesn't emerge because there are numerous dimensions for calculating costs and effectiveness: for example, short-run and long-run costs without agreement on a discount rate; initial and survivable capabilities; surge and sustainable capabilities; political impacts on allies or adversaries; etc.

Moreover, the various dimensions are likely to have different degrees of uncertainty and different weights attached to them by different groups outside as well as inside the policy community. Under these circumstances, policy analytic studies should, and sometimes do, display separately the various dimensions of cost and effectiveness, scoring the competing alternatives accordingly, and leaving choice to a subsequent decisionmaker or a decisionmaking process.

Even the most sophisticated analyses usually ignore or give meager attention to implementation issues. Defense planning studies rarely raise and almost never answer, such questions as who would have to what, and when, and with what possible and likely resistances, modifications, and compromises, if alternative A were chosen, or B, or C? It is therefore implicitly assumed that the costs and benefits as modeled in the analysis, won't be altered by implementation.

When this implicit prediction is made explicit, it will be readily acknowledged to be unwarranted, as is suggested by a vast range of cases; for example, development of the FB-111, and innumerable other instances of "goldplating" in the development of new weapons systems. The question rises whether we can do a better job in the future than we have in the past in systematically including implementation risks and prospects in the studies that we do?

If we are to answer this question in the affirmative, the part of the typical defense planning study that has been aptly named "the missing chapter," dealing systematically with implementation prospects, must become a standard part of defense policy studies.

In recent years, discussion of implementation issues has increased substantially. It has been concentrated in the new public policy journals, several recent books and case studies, and the curricula of graduate schools of policy analysis. Most of this discussion has emphasized the typically large gap between programs as designed and as executed, the lack of appropriate methods for anticipating these gaps and taking them into account in doing policy studies, and consequently the marked shortcomings of all defense planning analysis in failing to address implementation explicitly and systematically.

I have tried to deal with this set of issues elsewhere.⁶ In any event it would take me too far afield to try to summarize that discussion here. However, in conclusion, I predict that resource allocation and defense planning studies that are done in the future will and should devote much greater attention to implementation considerations than they have in the past.

6. Charles Wolf, "A Theory of 'Non-Market Failure': Framework for Implementation Analysis," *The Journal of Law and Economics*, April 1979.

THE CHALLENGE OF OPERATIONS RESEARCH
IN THE DEVELOPING COUNTRY

SANG M. LEE

University of Nebraska
Department of Management
Lincoln, Nebraska 68588

ABSTRACT. In this paper, the challenge of operations research in the developing country is discussed. There are three formidable obstacles that the operations research profession must face in the developing country: (1) the culture; (2) uncertainties of the decision environment; and (3) the lack of technical support resources. Overcoming these obstacles is a real challenge. However, there exist valuable sources of information that can be effectively used in facing this challenge: the scientific inquiry system and the history of operation's research in the industrialized nations. If such information is digested and adapted with creativity, it appears that operations research has an enormous potential for contribution toward helping the nation to achieve a sustained rate of growth in economic, social, and political frontiers of the developing country.

1. INTRODUCTION

Managerial decision problems have greatly increased in number, complexity, and magnitude over the past thirty years. This change has been due, in large part, to a dramatic change in the nature of management and the environment within which it operates. While organizations have become much larger and more complex, management functions have become more specialized. Natural resources required for operations have become increasingly scarce, forcing managers to evaluate decision problems while considering environmental constraints. Ecology and pollution control have become household words, governmental and consumer groups have begun to exert greater demands on organizational actions, and international politics and trade decisions (e.g., tension in the Middle East, the OPEC actions, etc.) have profound impact on many organizations.

As managers have become increasingly held accountable for their decisions, not only to top management and stockholders, but also to government and other outside interest groups, they have looked for more sophisticated approaches to the analysis of overwhelmingly complex problems. Thus there has been an increasing demand for techniques that would be helpful for finding the best solution to a problem and a defensible basis for arriving at the course of action selected.

Consequently, we have seen great progress in the field of operations research (OR), especially in the United States. Many new theories have been developed and existing techniques greatly refined as a result of technical breakthroughs. Also, new applications of existing techniques have been developed, and an increasing number of complex problems are being solved with the aid of the computer. The greatest advance in operations research, however, has occurred in the implementation of scientific approaches to real-world problems [10].

The students and practitioners of operations research in the developing country are faced with three formidable obstacles: (1) the culture which may not be conducive to the systematic decision making; (2) the uncertainties and changing nature of the decision environment; and (3) the lack of technical support resources. However, they also have a decisive advantage. They use the history of successes and failures of operations research in the industrialized nations as a guide. Nevertheless, there is an enormous challenge for operations researchers in the developing country. In this paper, the nature of this challenge will be discussed in light of the broad scientific inquiry system for decision making and the history of OR in the industrialized nations.

2. SCIENTIFIC INQUIRY AND DECISION MAKING

"A well-known scientist decided that he had been a bachelor long enough, or at least that he should seriously consider whether to get married or not, and if so to whom. Being a rational man, he sat down and enumerated the advantages and disadvantages of the marital state and the kind of qualities that he should look for in choosing a wife. As for the advantages--and I quote from his notes, 'Children (if it please God)--constant companion (and friend in old age)--charms of music and female chit-chat.' Among the disadvantages: 'Terrible loss of time, if many children forced to gain one's bread; fighting about no society.' But he continued, 'What is the use of working without sympathy from near and dear friends? Who are near and dear friends to the old, except relatives?' And his conclusion was: 'My God; it is intolerable to think of spending one's whole life like a neuter bee, working, working, and nothing after all.--No, no won't do.--Imagine living all one's day solitarily in smoky, dirty London house--only picture to yourself a nice soft wife on a sofa, with good fire and books and music perhaps--compare this vision with the dingy reality of Gt. Marlboro Street.' His conclusion: 'Marry, marry, marry.' Having decided that he ought to get married and having listed the desirable qualities of a future spouse, he then proceeded to look for a suitable candidate. He had several female cousins, so that there was no need to search outside the family circle. He dispassionately compared their attributes with his list of objectives and constraints, made his choice and proposed to her. Needless to say, he lived happily ever after. The scientist in question--Charles Darwin; the year--1837 [7]."

The above quotation is presented to point out several important aspects of scientific decision making process. The first aspect is that the rational decision making process is nothing new. Man's desire to take the most effective action has led to a continuous struggle to comprehend the norms and conditions under which the environment functions. The increase in man's body of knowledge has led to new discoveries, inventions, and innovations, and these have resulted in greater benefits, comforts, and challenges for man. The progress of a society has, therefore, been basically determined by the production and rate of increase in knowledge. In as much as knowledge is sought for more effective action, this pursuit is the basis for a cognitive system that we shall call "science." Science is the process or organized body of knowledge conforming to established rules of inquiry, as well as to a system of propositions [10].

Through knowledge, science, and specialization, man has recognized many of the relationships of his environmental systems. This new knowledge has provided man with the opportunity to manipulate environmental conditions to produce desired consequences. Man has thus acquired much control over nature, providing new horizons of civilization and growth. However, this is also the genesis of the problem of decision analysis, since different decisions may result in different consequences. Man attempts to select the course of action, from a set of alternative courses of action, that will achieve his objectives as fully as desired. Decision analysis is a formalized process for increasing man's understanding and control over environmental conditions. Thus, every new development in knowledge or science may have a potentially practical implication for decision analysis.

The second aspect we would like to point out from the quotation is that decision analysis is constrained by the environmental factors. Charles Darwin was a superb scientist and therefore he was able to be rational in selecting his course of action. For some reason, however, Darwin limited his search for the bride within the family circle. Although he thought he made the optimum decision, it might have been a suboptimum decision at best. This special constraint he imposed might have been due to his family training, personality, or the accepted life style during that period of time in England. In other words, the decision environment presents a host of constraints to the decision making process.

The third aspect which deserves our attention in the quotation is that Darwin employed a systematic methodology to solve a complex real-world problem that involves multiple objectives. It is precisely this element of complexity which has led to such a host of difficulties in scientific decision analysis.

2.1 Decision Environment

One of the primary incentives for man to pursue knowledge is the basic human and environmental problem of satisfying unlimited human desire with limited resources. This has always been the most troublesome human problem. Most human organizations, whether they are business enterprises, governmental agencies, or social institutions, have evolved in such a way as to narrow the gap between desires and resources.

There are two possible approaches that may be employed to solve this human problem. One is to increase available resources. For an individual, this approach may take the form of the Protestant work ethic which calls for hard work to increase resources so that the individual can satisfy

most of his desires. Or it may take a form of scientific endeavors that will enable him to utilize existing limited resources in a more efficient manner. Finally, it may mean a scientific breakthrough that creates new uses for relatively abundant resources, such as water, air, sunshine, etc.

The second basic approach is for an individual to limit his desires so that the existing resources become sufficient to satisfy them. For example, one may choose to have a small cottage in the mountains and meditate ten hours, have only two meals, and work four hours per day. The two approaches we have cited are quite in contrast, but both have found wide practice in the history of human society [3].

The first approach is clearly a general philosophy of Western culture. One who accumulates wealth through hard work or innovative ideas becomes a successful person. The economic rewards of hard work also usually result in social and psychological rewards. In short, "money talks" for the fulfillment of human desires. It is quite common in this country to find a millionaire being respected even more highly than statesmen, artists, scholars, or religious leaders. Since we live in an environment of scarce resources (and it is becoming scarcer everyday), it may be perfectly natural and appropriate for those who acquire greater control over resources to receive social respect.

The second approach has been a long-accepted practice, although it is gradually diminishing, in Eastern cultures. By exercising strong self-discipline, self-control, and sometimes even self-denial, one reduces his desires to the very minimal, let's say to the subsistence level. This philosophy is broadly defined as asceticism. Ascetics usually introduce some philosophic or religious accent into their daily life in order to enrich their "inner" happiness. Physiologically, the practice of asceticism is far from a comfortable way of life. However, many have found the practice of asceticism a meaningful life style as they receive social respect for their philosophy, knowledge, and courage to endure physical hardships. If not completely ascetic, an austere way of life has long been advocated in many Asiatic countries as the gentlemanly life style. For example, even today, many oriental millionaires spend their leisure by enjoying "small pleasures" at home, such as practicing calligraphy, playing "go" (complicated oriental chess-type game), composing poems, or watching the birds.

The point of this discussion, aside from the pros and cons of Eastern and Western culture, is that in a society where limitation of human desires is a respected way of life, it is unlikely that systematic decision analysis will find an important role to play. In other words, OR becomes important in a cultural setting where the way of life is geared toward greater control over resources to fulfill desires. Therefore,

decision science is not equally applicable to a given problem in different decision environments.

2.2 The Concept of Rationality

The traditional economic theory postulates an "economic man," who is "economic" and also "rational." The economic man is an "optimizer" in the Western cultural sense. He is assumed to be one who allocates his resources in the most rational manner and has the knowledge of the relevant aspects of his environment. He is also assumed to possess a stable system of preferences and the skill to analyze the alternative courses of action in order to achieve his desires. In the classical economic theory, we often assume that the concept of economic man provides the basic foundation for the theory of the firm. In other words, we often think that the decision-making process of an organization is or should be like one employed by the economic man.

Recent developments in the theory of the firm have cast considerable doubt on whether the concept of economic man can be applied to the decision maker in today's complex organizations [20]. According to broad empirical investigation, there is no evidence that any one individual is capable of performing a completely rational analysis for complex decision problems. Also, there is considerable doubt that the individual value system is exactly identical to that of the firm in determining what is best for the organization as a whole. Furthermore, the decision maker in reality is often quite incapable of identifying the optimum choice, because of either his lack of analytical ability or the complexity of the organizational environment. The concept of economic man does not sufficiently provide either a descriptive or a normative model for the decision maker in an organization. Because of the organismic limitations of the decision maker, his decision making will at best be a crude approximation of global rationality. In this context, a well-known management theorist professor H. A. Simon (1978 Nobel Laureate), has suggested that in today's complex organizational environment, the decision maker is not trying to optimize, instead he tries to satisfice [21].

There is an abundance of evidence that suggests that the practice of decision making is affected by the epistemological assumptions of the individual who makes the decision. Indeed, the practice of scientific methodology and rational choice are not always directly applicable to decision analysis. The decision maker is then, in reality, one who attempts to employ an "approximate" rationality in order to maximize the attainment of organizational goals within the given set of constraints. He may fall far short of being a completely

rational man, but his decision making behavior may at least be "intentionally" rational, or he has the "bounded" rationality. If we define decision science as a rational choice process within the context of the decision maker's environmental concern and his limited knowledge, ability, and information, the paradox between the economic man and decision maker in reality becomes increasingly vague [21]. There still remains discrepancies between the theory of rationality and realities of human life. These discrepancies, however, may provide valuable information for the analysis of human behavior in the organizational environment.

2.3 Multiple Objectives

Organizational objectives vary according to the character, type, philosophy of management, and particular environmental conditions of the organization. There is no single universal goal for all organizations. Profit maximization, which is regarded as the sole objective of the business firm in the classical economic theory, is one of the most widely accepted objectives of management. As reviewed above, in today's dynamic business environment, profit maximization is not always the only objective of management. In fact, business firms quite frequently place higher priorities on non-economic goals than on profit maximization. Or, firms often seek profit maximization while pursuing other non-economic objectives. We have seen, for example, firms place a great emphasis on social responsibilities, social contributions, public relations, industrial and labor relations, etc. Whether such objects are sought because of outside pressure or voluntary management decisions, non-economic objectives exist and they are gaining a greater significance. The recent public awareness of the need for ecology management and the gaining momentum of consumerism may have forced many firms to reevaluate their organizational objectives. An exhaustive study by Lee [10], and Shubik [20] clearly indicates that firms strive to fulfill multiple goals.

Many contemporary decision problems faced by industry, government, and other institutions will increasingly require identification of more elusive and abstract objective functions. The objective function no longer will be restricted to a cardinal criterion; rather it will involve general criteria related to the common good. Certainly, costs will remain to be important decision variable because it determines the resource requirements. However, its function will be shifted from that of the objective function to a decision constraints.

Important developments in the field of management have clearly indicated that "management by multiple objectives"

is the most difficult and important area of operations research. Martin K. Starr, editor of Management Science and a past president of the Institute of Management Science (TIMS), stated at a recent professional conference that in his opinion the most important research topic in the field of operations research today is the area of multiple criteria decision analysis. Warren Bennis, an eminent scholar in the field of organizational development, stated at a professional conference that organizations, as well as society in general, have become so fragmented into various interest and value groups that there is no longer one predominant objective for any organization. Consequently, one of the most important and difficult aspects of any decision problem is to achieve an equilibrium among the multiple and conflicting interests and objectives of the various components of the organization. Several recent studies concerning the future of the industrialized society [19] have echoed the same theme. When the society is based on enormous technological development and change, stability of the system must be obtained by achieving a delicate balance among such multiple objectives as food production, industrial output, pollution control, population growth control, use of nonrenewable natural resources, and international cooperation for economic stability.

One of the most promising operations research techniques for multiple objective decision analysis is goal programming. Goal programming is a powerful tool which draws upon the highly developed and tested technique of linear programming, but provides a simultaneous solution to a complex system of competing objectives. Goal programming can handle decision problems having a single goal with multiple subgoals, as well as cases having multiple goals and subgoals [10]. The concept of goal programming was originally introduced by Charnes and Cooper [1], [2], [4], and further studied by Ijiri [9], Lee [10], [11], [12], and others [8]. Application of goal programming to real-world decision problems have been explored for advertising media planning [6], manpower planning [5], production planning [15], academic planning [13], financial decision making [14], economic policy analysis [10], transportation logistics [16], [17], marketing strategy planning [18], environmental protection [3], health care planning [10], and many others.

3. OPERATIONS RESEARCH IN THE DEVELOPING COUNTRY

Many developing nations have achieved truly miraculous levels of economic growth during the past 10 to 15 years. We can witness such growth and advancement here in Korea. Perhaps we can single out the following factors as the most important ingredients of the impressive economic growth

achieved by several developing countries: a strong sense of common purpose for survival and prosperity among people; determination and a high level of motivation for hardwork by workers; careful economic planning and vigorous international trade; and the quality of leadership in government and business. Although these ingredients are important for the economic growth, they are not sufficient conditions for a stable and continuous rate of growth. The primary reason for this is the fact that the expanded base of the economy, life style, and technological requirements require additional stimuli for a continuous growth.

When the developing country reaches a point in her economic growth where it becomes an effective competitor with industrialized nations, many existing internal and external constraints become increasingly more burdensome. For example, the following major obstacles must be successfully succumbed in order to achieve a stable rate of economic growth: stable supply sources of required raw materials (especially, from external sources); international and domestic markets for new products; national security and political stability of the nation; timely and necessary governmental policy changes for national and international business; political and economic pressures in the international scene; technological assistance needs; and cultural barriers for systematic growth of the economy.

Operations research can be utilized as a universal tool in alleviating many of the obstacles discussed above. For example, OR can be utilized for the national defense planning, development of a comprehensive economic planning models, forecasting changes in the international trade; and the like. As a matter of fact, operations research can serve as a vehicle that can greatly improve the conceptual skills of the management manpower in the developing country. It is a rule rather than exception that most developing countries emphasize technical skills (e.g., technical high schools and colleges, skill-oriented vocational training, engineering and drafting skills education, etc.) and neglect management skills (e.g., formal training for leadership, motivation, analytical and conceptual skills, and decision making). It is analogous to emphasizing the marksmanship of the soldiers and neglecting tactics and strategies in the military.

The operations research profession in the developing country has difficult yet challenging roles to play. It can help the nation to achieve not only a sustained rate of economic growth and the standard of living, but it can also serve as the pole-bearer for systematic analysis and improved decision making on the part of the managers and administrators. It is hoped that operations research would become a core requirement for business and administration programs at colleges and universities. This certainly would speed up

the understanding and actual use of operations research in the developing country.

4. THE FUTURE OF OPERATIONS RESEARCH IN THE DEVELOPING COUNTRY

As discussed earlier, the developing country has certain advantages over the industrialized nation in applying operations research. It can utilize the experience of others' successes and try to avoid their failures. Also, it can easily start the use of operations research with the most advanced technical (theories and techniques) and technological (computer hardwares and software) tools available today rather than going back to 25 years ago. Let us discuss several approaches that can be explored by operations researchers in the developing country.

4.1 Learning from the Failures of Others

The development stages of operations research in the United States can be broadly classified as follows: (1) the primitive stage (prior to 1960's); (2) the rapid growth stage (the 1960's); and (3) the maturing stage (the 1970's). During the primitive stage, the operations research profession was organized (e.g., TIMS and ORSA) by those who transferred from such disciplines as mathematics, statistics, natural sciences, and engineering. Operations researchers were primarily interested in learning and developing techniques in order to find optimum solutions to clearly defined operational problems, such as production scheduling, inventory problems, blending problems, and the like.

The rapid growth stage was the period during which a dramatic growth of operations research occurred in academic institutions. The rapid growth of OR in academic institutions has brought some positive results. Perhaps the most important development was that OR provided a special impetus to utilize the enormous analytical power of computers to decision analysis. However, there also were many negative results. Many operations researchers have put a great deal of energy into academic, theoretical, or pure research for refinement of minute details of various techniques that had no or little relevance to real-world problems. Or, they tried to apply powerful tools to tedious and unimportant problems. Consequently, research was often conducted for the sake of research or publication. Many practitioners and managers began to be disillusioned, partially due to their inability to comprehend the research work and partially because of the irrelevance of the study. Also, evidence of a tendency began to develop which emphasized the techniques

over the problems to be solved and some researchers tended to look for problems that could be simplified and solved by certain techniques. This "have-gun-will-travel" approach is indeed contradictory to the very purpose of OR. There has been a lack of understanding of the decision environment, organizational values, conflicting nature of objectives, data requirements, time constraints, politics of the organization, and noneconomic ramifications of decision alternatives.

During the maturing stage, the OR profession began a self-evaluation. The operations researchers started to ask the question, "Are we doing the job we are supposed to be doing?" Many leading OR practitioners began to speak out about the failures of OR as well as its successes. This phase can be characterized by its emphases on: more pragmatic approaches that would ensure actual implementation of OR studies; analysis of the entire decision environment and the nature of the problem rather than simply developing a model; obtaining satisficing solution rather than always looking for the optimal solution; application of advanced computer technology (simulation, heuristics, etc) for ill-structured managerial problems; and multiple objective decision making approaches.

The OR profession in the developing country should not experience the same developmental pains as in the United States. It should avoid the failures experienced by OR practitioners in the United States, especially the aimless emphasis of theories, the tendency to overkill a problem with sophisticated techniques, and the OR travelling salesman approach. It should emphasize the aspects of OR that are currently important in the developing country, such as: developing simple and cost-effective approaches to problems; analyzing a problem in its entire environmental perspective; developing a satisficing schemes rather than always looking for the optimal solution which may not even exist; emphasizing the implementation of OR models rather than simply formulating models in order to show off the researcher's mathematical prowess; and recognizing the multiple and often conflicting objectives in every important decision problem.

4.2 Learning from the Successes of Others

The OR profession in the developing country can also benefit from the success experiences of OR in the United States. The following items probably represent the most important aspects of the OR success stories.

4.2.1 Interdisciplinary Approach

In OR, problem solving is usually approached by a team

rather than by one researcher. There are several reasons for this interdisciplinary (or at least a task-force) approach. First, the body of knowledge in OR is so vast that it becomes impossible for one individual to be a jack-of-all-trades. For example, a researcher may be an established queueing expert, but he may not know enough linear programming to develop a production planning model. Second, the problem in question may involve several departments in the organization. The researcher may have neither the organizational experience nor the time to comprehend all the interrelationships among the departments. A convenient way to acquire this knowledge and also have the operating personnel participate in the decision analysis is to form a team by selecting a key person from each department. Third, many complex problems usually involve economic, biological, psychological, physical, engineering, and environmental aspects. These aspects of the problem can be analyzed best by those who are directly involved in these areas.

4.2.2. Strategies of OR Implementation

For a successful implementation of OR studies, a considerable amount of artistic creativity is required. The real-world application of OR requires creativity to combine OR knowledge with the complexities of the problem. Furthermore, the researcher must believe in the value of his work. Without this confidence, he cannot effectively communicate his ideas and expertise to the manager. The application of OR requires information such as management philosophy, policies, goals, and relationships among pertinent decision variables. Much of such information can be obtained only from top management. The researcher must have the full confidence of the manager for problem analysis and implementation.

There have been several successful attempts in the United States to narrow the gap between management and OR specialists. First, there has been a rapid decentralization of OR programs in industrial organizations. Instead of maintaining a separate OR group, many firms have sprinkled OR specialists throughout the organization in positions where they can really help the organization. Many OR specialists are assigned to significant functional positions, or as aides to the top managers. They are given line responsibilities for results. Another trend is the manager's demand for the decision model implementation by the OR group. This approach has been successful in alleviating the problem of unworkable, theoretical model design on the part of the OR group. A third trend has been a thorough on-the-job training of OR specialists about the functional, social, behavioral, and

political aspects of the decision environment. It is the responsibility of management to expose OR specialists to the real-world decision environment. In summary, everyone recognizes the potential contribution of OR. The issue that needs our attention is the best way to put OR to work.

4.2.3 Multiple Objective Decision Making

As discussed earlier, multiple objective decision analysis has become a very important area of OR in recent years. It is a challenge to the OR profession in the developing country to apply this concept to intertwined decision problems they now face. Achievement of various economic and social goals while maintaining heavy national defense and security burdens clearly presents a multiple objective decision problem. Analysis of the problem in terms of appropriate goal levels, priorities, trade-offs, sensitivity of the solution, and various social, economic, and political ramifications of solutions would provide very valuable information. Many new advances in the area of goal programming (e.g., interactive, integer, decomposition, separable, and chance-constrained approaches) may be valuable tools for the analysis of multiple conflicting objective decision problems.

4.2.4 Technical Progress

In order to analyze the ever-increasing complexity of managerial problems, more sophisticated OR techniques are needed. Indeed, there have been many important advances made during the past 10 years, such as simulation techniques, heuristics, multiple objective models, etc. The second advance we have seen is in the expanded use of existing techniques. For example, there has been such an improvement in the efficiency of linear programming that it can be applied to very large-scale problems involving thousands of decision variables and constraints. A continuous development of new and better theories and expansion of existing techniques would be necessary to analyze many contemporary societal problems. A third area of advance in OR has been the development of more realistic descriptive models for managerial problems. Up until now, OR specialists have primarily engaged in the development of normative-type models. In other words, the model has been developed to get an answer as to how a decision ought to be made. However, since the model has often been designed without the consideration of the decision makers' philosophy and other environmental factors, the model result has found only occasional implementation. The descriptive model incorporates many of the important environmental aspects and the decision makers'

judgment. Such a model would undoubtedly be more acceptable for implementation.

The OR profession in the developing country must realize the fact that the technical progress is a very costly proposition. Thus, it should develop a cooperative scheme with either leading universities or professional organizations in the United States and/or other advanced nations so that most up-to-date information can be easily obtained at reasonable costs.

4.2.5 Technological Progress

The technical progress provides the necessary means to perform a systematic decision analysis. Actual applications of OR to complex real-world problems, however, require the use of the computer. The remarkable progress in computer technology, both in hardware and software areas, has allowed a greater sophistication of decision models. It is now possible to construct and solve very elaborate production, inventory, finance, and planning models that require a great memory capacity on the part of the computer. With the continuing inventions and innovations in the computer field, the application of OR is expected to become even a more important management function.

Already, micro computers and the time-sharing method have brought dramatic changes in decision analysis. With the convenience of a remote terminal facility, a continuous monitoring of a decision system on a real time basis is possible. With the progress we have seen in computer technology, many repetitive type operating systems can be analyzed entirely by the computer. Another important advance in computer technology that has had a significant impact on the application of OR is the standardization of various techniques in the form of "software packages." With further advances in this area, perhaps the application of the most widely used OR techniques may become routine.

The OR profession in the developing country must devise a scheme to obtain information concerning the most up-to-date computer hardware and softwares. Evaluation of the available computer resources and selecting the most appropriate systems would be a very important role of OR specialists.

4.2.6 Organizational Impact

In the beginning of the OR application in various organizations, an operations researcher was an isolated person in the organization. He designed mathematical models whenever the need arose but remained isolated from the actual implementation and impact of the study. However, today, management personnel recognize the impact of OR studies on the

organization as well as on the work behavior of those directly affected by them.

Operations research specialists are professionals with advanced degrees. They are usually specialists with narrow backgrounds. In order to play their role effectively, they must learn the structure, functions, work methods, and behavioral aspects of the organization. Furthermore, they must be patient persuaders and advisers. In other words, the effective operations researcher must be "a specialist with a universal mind."

The OR specialist in the developing country must be cognizant to the pragmatic aspects of the organization. They should recognize the importance of the external factors, internal environment of the organization, and behavioral implications of OR studies. The impact of OR studies would most likely to be very dramatic and profound in the developing country if operations researchers are able to win over the culture, management, and the system to their side.

5. CONCLUSION

In this paper the roles and challenges of OR in the developing country are explored. There are many obstacles that must be overcome by OR specialists such as the culture which is not conducive to the systematic decision analysis, uncertainties of the decision environment, overwhelming priority of the national security, and the lack of technical and technological support resources. However, this is precisely why the challenge of operations research provides excitement and sense of special pride to the operations researchers. This paper attempts to provide positive notes as to how this challenge can be met. It presents the basics of the scientific inquiry system and the history of operations research in the United States as possible sources of information. If this information is effectively utilized, the potential contribution of operations research in the developing country is indeed unlimited.

REFERENCES

- [1] Charnes, A., and W. W. Cooper, MANAGEMENT MODELS AND INDUSTRIAL APPLICATIONS OF LINEAR PROGRAMMING, Wiley, New York, 1961, 2 vols.
- [2] _____, _____, and R. O. Fugerson, OPTIMAL ESTIMATION OF EXECUTIVE COMPENSATION BY LINEAR PROGRAMMING, Management Science, 1 NO 2, 1955

- [3] _____, J. Harrauld, K. Karwan, and W. A. Wallace, A GOAL INTERVAL PROGRAMMING MODEL FOR RESOURCE ALLOCATION IN A MARINE ENVIRONMENTAL PROTECTION PROGRAM, Rensselaer Polytechnic Institute, School of Management Research Report, September, 1975.
- [4] _____, D. Klingman, and R. J. Niehaus, EXPLICIT SOLUTIONS IN CONVEX GOAL PROGRAMMING, Management Science, 22, NO 4, 1975
- [5] _____, and R. J. Niehaus, STUDIES IN MANPOWER TRAINING, U.S. Navy Office of Civilian Manpower Management, Washington, 1972
- [6] _____, et al., NOTE ON THE APPLICATION OF A GOAL PLANNING MODEL FOR MEDIA PLANNING, Management Science, 14, NO 8, 1968
- [7] Eilon, S., GOALS AND CONSTRAINTS IN DECISION-MAKING, Operational Research Quarterly, XXIII, NO 1, 1973, PP. 3-15
- [8] Ignizio, J. P., GOAL PROGRAMMING AND EXTENSIONS, Lexington Books, Lexington, Mass., 1976
- [9] Ijiri, Y., MANAGEMENT GOALS AND ACCOUNTING FOR CONTROL, Rand-McNally, Chicago, 1956
- [10] Lee, S. M., GOAL PROGRAMMING FOR DECISION ANALYSIS, Auerbach, Philadelphia, 1972
- [11] _____, GOAL PROGRAMMING FOR DECISION ANALYSIS, WITH MULTIPLE OBJECTIVES, Sloan Management Review, 14, NO 2, 1973
- [12] _____, INTERACTIVE AND INTEGER GOAL PROGRAMMING, Paper presented at the Joint ORSA/TIMS Meeting, Las Vegas, 1975
- [13] _____, and E. R. Clayton, A GOAL PROGRAMMING MODEL FOR ACADEMIC RESOURCE ALLOCATION, Management Science, 17, NO 8, 1972
- [14] _____, and A. J. Lerro, OPTIMIZING THE PORTFOLIO SELECTION FOR MUTUAL FUNDS, Journal of Finance, 28, NO 8, 1972
- [15] _____, and L. J. Moore, A PRACTICAL APPROACH TO PRODUCTION SCHEDULING, Production and Inventory Management, 15, NO 1, 1974

- [16] _____, and _____, OPTIMIZING TRANSPORTATION PROBLEMS WITH MULTIPLE OBJECTIVES, AIIE Transactions, 5, NO 4, 1973
- [17] _____, and _____, MULTI-CRITERIA SCHOOL BUSING MODELS, Management Science, Vol 23, NO 7, PP. 703-715, 1977
- [18] _____, and . Nicely, GOAL PROGRAMMING FOR MARKETING DECISIONS: A CASE STUDY, Journal of Marketing, 38, NO 1, 1974
- [19] Meadows, D. H., et al., THE LIMITS TO GROWTH, The New American Library, Washington, D.C., 1972
- [20] Schubik, M., APPROACHES TO THE STUDY OF DECISION-MAKING RELEVANT TO THE FIRM, in The Making of Decisions: A Reader in Administrative Behavior, ed., W. J. Gore and J. W. Dyson, The Free Press of Glencoe, London, 1964
- [21] Simon, H. A., THE NEW SCIENCE OF MANAGEMENT DECISION, Harper & Brothers, New York, 1960

SOME NEW MODELS OF
QUEUEING THEORY

TOSHIO NISHIDA

Dept. of Applied Physics
Faculty of Engineering
Osaka University
Osaka, Japan

1. INTRODUCTION.

A queueing system exists when a customer arrives at a service facility, wait in queue for service, are serviced, and then depart. There is a number of variations in each of the phenomena associated with a queueing system. We have to describe these phenomena in a mathematically manipulatable form.

It is well-known that the origin of queueing theory is the work by A.K. Erlang in 1910 concerning telephone congestion. Thus, this theory has about 70 years history. It has been applied to a great variety of business situations. For example, checkout stations of supermarkets, restaurants, gasoline stations, airline counters, hospitals, production lines, and so on.

On the other hand, a huge number of theoretical papers has been published. There are about 9 hundred papers in the references of the famous book "Elements of queueing theory" by T.L. Saaty.

Table 1 shows the number of papers in recent 5 years concerning queueing theory which is taken from the International Abstracts in Operations Research.

In this paper, I shall present two new models of queueing theory. One is correlated multi-server queueing model, and the other is commutative tandem queues.

TABLE 1

	1974	1975	1976	1977	1978
Theory	61	37	34	53	50
Applications	19	21	24	40	37
communication			1	2	2
computer	2	2	2	3	5
distribution				2	
education			1		
gaming	1				
health service	1		2	5	3
inventory			1	2	3
location	1				
maintenance	2	2	1	3	3
production	5	1	5	4	3
public service		4	4	2	1
scheduling	1	2	2	3	
simulation	1	3		1	1
transportation	5	7	5	13	16
total	80	58	58	93	87

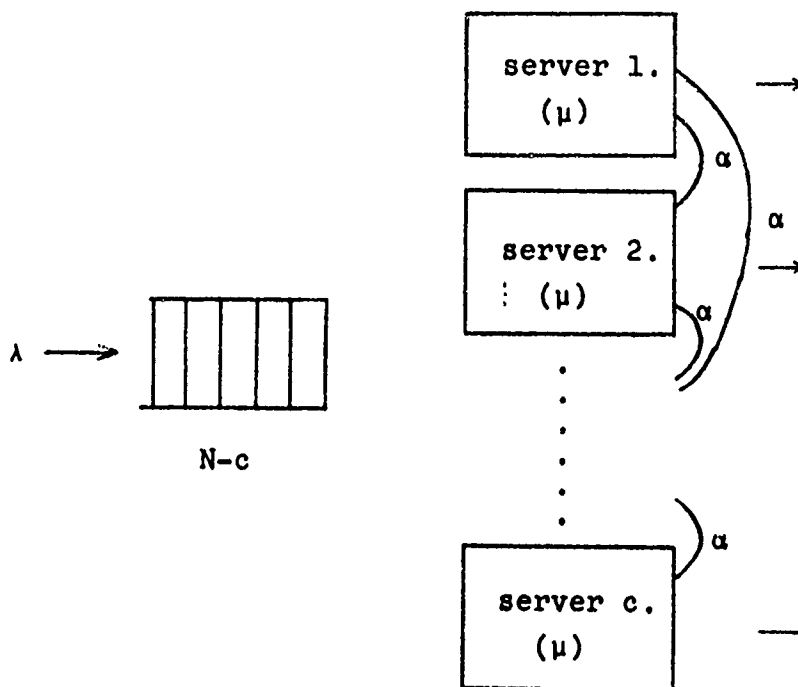
2. CORRELATED MULTISERVER QUEUE

In most of the published studies of multiserver queuing systems, it has been assumed that the service time of all servers are mutually independent. In some practical situation, however, this assumption is not realistic, due to the competition or cooperation among servers.

2.1 General Model with Random Input

We shall consider a multiserver queuing model which has arbitrary c servers. It is assumed that the successive customers arrive according to a Poisson stream with a parameter $\lambda > 0$, and the system is able to hold a maximum of N customers, and the ordinary queue discipline is being employed. The service time of c servers are assumed to follow the modified form of multivariate exponential distribution (MVE) introduced by Marshall and Olkin, where the service rate at which only one service finishes is all equal to μ , the service rate at which two services finish at the same time is all equal to α and the event equal to or more than three services finish at the same time cannot occur since it is not realistic in the queuing model. We shall denote this system by the symbol $M / MVE / c(N)$.

Fig.1. $M/MVE/c(N)$ queuing system



Let x_i ($1 \leq i \leq c$) be random variables representing the service time of c servers. Then, by the assumption, we have

$$\begin{aligned} & \Pr [X_1 > x_1, X_2 > x_2, \dots, X_c > x_c] \\ &= \exp \left[-\mu \sum_{i=1}^c x_i - \alpha \sum_{i=1}^c \sum_{j=i+1}^{c-1} \max(x_i, x_j) \right]. \end{aligned} \quad (1)$$

From (1), γ -dimensional marginal distribution is

$$\exp \left[-[\mu + (c-r)\alpha] \sum_{i=1}^r x_{k_i} - \alpha \sum_{i=1}^r \sum_{j=i+1}^{r-1} \max(x_{k_i}, x_{k_j}) \right] \quad (2)$$

especially,

$$\begin{aligned} & \Pr [\min(X_1, X_2, \dots, X_r) > x] \\ &= \exp \left[-\frac{1}{2} r [2\mu + (2c-r-1)\alpha] x \right] \end{aligned} \quad (3)$$

The fact that this distribution is exponential will be useful in the following.

Let $P_n(t)$ denote the probability that the number of customers in the system is n at time t . Then we can write the differential-difference equations for this system as follows:

$$\begin{aligned} P'_0(t) &= -\lambda P_0(t) + a_1 P_1(t) + b_2 P_2(t), \\ P'_n(t) &= \lambda P_{n-1}(t) - (\lambda + a_n) P_n(t) + (a_{n+1} - b_{n+1}) P_{n+1}(t) + b_{n+2} P_{n+2}(t), \\ &\quad (1 \leq n \leq c-2) \\ P'_{c-1}(t) &= \lambda P_{c-2}(t) - (\lambda + a_{c-1}) P_{c-1}(t) + (a_c - b_c) P_c(t) + b_c P_{c+1}(t), \\ P'_n(t) &= \lambda P_{n-1}(t) - (\lambda + a_c) P_n(t) + (a_c - b_c) P_{n+1}(t) + b_c P_{n+2}(t), \\ &\quad (c \leq n \leq N-2) \\ P'_{N-1}(t) &= \lambda P_{N-2}(t) - (\lambda + a_c) P_{N-1}(t) + (a_c - b_c) P_N(t), \\ P'_N(t) &= \lambda P_{N-1}(t) - a_c P_N(t), \end{aligned} \quad (4)$$

where

$$a_n = \frac{1}{2} n [2\mu + (2c-n-1)\alpha], \quad b_n = \frac{1}{2} n(n-1)\alpha, \quad 1 \leq n \leq c. \quad (5)$$

From these equations, we can readily obtain the following difference equations for the steady-state probabilities

$$p_n = \lim_{t \rightarrow \infty} P_n(t)$$

$$\begin{aligned} -\lambda p_n + a_{n+1} p_{n+1} + b_{n+2} p_{n+2} &= 0, & (0 \leq n \leq c-2) \\ -\lambda p_n + a_c p_{n+1} + b_c p_{n+2} &= 0, & (c-1 \leq n \leq N-2) \\ -\lambda p_{N-1} + a_c p_N &= 0. \end{aligned} \quad (6)$$

And the normalization condition is

$$\sum_{n=0}^N p_n = 1. \quad (7)$$

In order to solve (6) and (7), we use the generating function

$$F(z) = \sum_{n=0}^N p_n z^n. \quad (8)$$

From the equations in (6) except for the first $(c-2)$ equations, we get

$$\begin{aligned} F(z) &= [\lambda p_N z^{N+2} - (\mu z + b_c) z^{c-1} p_{c-1} - (a_c z + b_c) z^{c-2} p_{c-2} \\ &\quad + f(z) \sum_{n=0}^{c-3} z^n p_n] / [f(z)], \end{aligned} \quad (9)$$

where

$$f(z) = \lambda z^2 - a_c z - b_c. \quad (10)$$

Normalization condition yields

$$-\lambda p_N + (\mu + b_c) p_{c-1} + (a_c + b_c) p_{c-2} + (a_c + b_c - \lambda) \sum_{n=0}^{c-3} p_n = a_c + b_c - \lambda \quad (11)$$

And the regularity of $F(z)$ shows

$$\begin{aligned} \lambda p_N \xi^{N-c+3} - (\mu \xi + b_c) p_{c-1} - \lambda \xi p_{c-2} &= 0 \\ \lambda p_N \xi^{N-c+3} - (\mu \xi + b_c) p_{c-1} - \lambda \xi p_{c-2} &= 0 \end{aligned} \quad (12)$$

where

$$\xi = [a_c - (a_c^2 + 4\lambda b_c)^{\frac{1}{2}}] / (2\lambda), \quad \zeta = [a_c + (a_c^2 + 4\lambda b_c)^{\frac{1}{2}}] / (2\lambda) \quad (13)$$

are the zeros of (10).

To determine p_n ($0 \leq n \leq c-1$) and p_N , we can use (11), (12) and the first $(c-2)$ equations in (6). Remaining p_n ($c \leq n \leq N$) are

$$p_n = p_N (\xi^{N+1-n} - \zeta^{N+1-n}) / (\xi - \zeta), \quad (c \leq n \leq N-1). \quad (14)$$

If we set $N=c$, we can get the results for no queue case. For the case $N=\infty$, we need the steady-state condition

$$\lambda < c [\mu + (c-1)\alpha] \quad (15)$$

and the generating function becomes

$$G(z) = [-(\mu z + b_c)z^{c-1}p_{c-1} - (a_c z + b_c)z^{c-2}p_{c-2}] / [f(z)] + \sum_{n=0}^{c-3} z^n p_n. \quad (16)$$

2.2. Explicit Solutions for Two and Three Servers Cases

2.2.1. Three Servers Case (M/MVE/3(N))

In this case, the generating function is

$$\begin{aligned} F(z) &= \frac{\lambda p_N z^{N+2} - \mu p_2 z^3 - (2\alpha p_2 + 2\mu p_1 + \alpha p_1)z^2 - (3\alpha p_1 + 3\mu p_0 + 3\alpha p_0)z - 3\alpha p_0}{\lambda(z - \xi)(z - \zeta)} \end{aligned} \quad (17)$$

where

$$\xi = [3(\mu + \alpha) - [9(\mu + \alpha)^2 + 12\alpha\lambda]^{\frac{1}{2}}] / (2\lambda) \quad (18)$$

$$\zeta = [3(\mu + \alpha) + [9(\mu + \alpha)^2 + 12\alpha\lambda]^{\frac{1}{2}}] / (2\lambda).$$

The steady-state probabilities are

$$\begin{aligned} p_0 &= (6\alpha + 3\mu - \lambda)B_1/B_2, \\ p_1 &= \lambda [3\alpha(\xi^N - \zeta^N) + \mu\xi\zeta(\xi^{N-1} - \zeta^{N-1})] p_0/B_1, \\ p_n &= 3\alpha\lambda(\xi^{N+1-n} - \zeta^{N+1-n})p_0/B_1, \quad (2 \leq n \leq N). \end{aligned} \quad (19)$$

where

$$B_1 = 3\alpha(2\alpha + \mu)(\xi^N - \zeta^N) + \xi\zeta(\xi^{N-1} - \zeta^{N-1})(\mu^2 + 2\alpha\mu - \alpha\lambda) \quad (20)$$

$$B_2 = (2\alpha + \mu)[3\alpha(6\alpha + 3\mu + 2\lambda)(\xi^N - \zeta^N) + \xi\zeta(\xi^{N-1} - \zeta^{N-1})(6\alpha\mu + 3\mu^2 + 2\mu\lambda - 3\alpha\lambda - \lambda^2)] - 3\alpha\lambda^2(\xi - \zeta)$$

Upon setting $N = 3$, we have the following results for no queue case.

$$\begin{aligned} p_0 &= C_1/C_2, \\ p_1 &= 3\lambda(3\alpha^2 + 2\mu^2 + 5\alpha\mu + \alpha\lambda)p_0/C_1, \\ p_2 &= 3\lambda^2(\alpha + \mu)p_0/C_1, \\ p_3 &= \lambda^3 p_0/C_1, \end{aligned} \quad (21)$$

where

$$\begin{aligned} C_1 &= 3(2\alpha + \mu)(3\alpha^2 + 2\mu^2 + 5\alpha\mu + \alpha\lambda) + 3\alpha\lambda(\alpha + \mu), \\ C_2 &= 3(2\alpha + \mu + \lambda)(3\alpha^2 + 2\mu^2 + 5\alpha\mu + \alpha\lambda) + 3\lambda(\alpha + \mu)(\alpha + \lambda) + \lambda^3 \end{aligned} \quad (22)$$

And for the infinite queue case, under the condition $\lambda < 3(\mu + 2\alpha)$, letting $N \rightarrow \infty$, we obtain the following:

$$\begin{aligned} \frac{B_1}{B_2} &\rightarrow \frac{(\mu^2 + 2\alpha\mu - \alpha\lambda)\xi + 3\alpha(2\alpha + \mu)}{(2\alpha + \mu)[(6\alpha\mu + 3\mu^2 + 2\mu\lambda - 3\alpha\lambda - \lambda^2)\xi + 3\alpha(6\alpha + 3\mu + 2\lambda)]} \equiv \frac{D_1}{D_2} \\ \tilde{p}_0 &= \lim_{N \rightarrow \infty} p_0 = (6\alpha + 3\mu - \lambda)D_1/D_2, \\ \tilde{p}_1 &= \lim_{N \rightarrow \infty} p_1 = \lambda(3\alpha + \mu\xi)\tilde{p}_0/D_1, \\ \tilde{p}_n &= \lim_{N \rightarrow \infty} p_n = 3\alpha\lambda\xi^{1-n}\tilde{p}_0/D_1, \quad (n \geq 2) \end{aligned} \quad (23)$$

If $\alpha = 0$, we have

$$\begin{aligned} \hat{p}_0 &= \lim_{\alpha \rightarrow 0} p_0 = \frac{2(1 - \rho)}{2 + 4\rho + 3\rho^2 - 9\rho^{N+1}} \\ \lim_{\alpha \rightarrow 0} p_1 &= 3\rho\hat{p}_0, \\ \lim_{\alpha \rightarrow 0} p_n &= \frac{9}{2}\rho^n\hat{p}_0, \quad (2 \leq n \leq N) \end{aligned} \quad (24)$$

which agrees with the solution of M/M/3(N) queuing system.

2.2.2. Two Servers Case (M/MVE/2(N))

In this case, we require only equations (11) and (12). We use, in this subsection, ν instead of α , α and β instead of ξ and ζ respectively. Then, we have

$$\begin{aligned}\alpha &= [(2\mu + \nu) - \sqrt{(2\mu + \nu)^2 + 4\lambda\nu}] / 2\lambda \\ \beta &= [(2\mu + \nu) + \sqrt{(2\mu + \nu)^2 + 4\lambda\nu}] / 2\lambda\end{aligned}\quad (25)$$

The steady-state probabilities are

$$\begin{aligned}p_0 &= \frac{2(\mu + \nu) - \lambda}{2(\mu + \nu) + (\mu + \nu)A_1/A_0 - \lambda A_N/A_0} \\ p_n &= (A_n/A_0)p_0, \quad n=0, 1, \dots, N.\end{aligned}\quad (26)$$

where

$$\begin{aligned}A_0 &= (\mu\alpha + \nu)\beta^{N+1} - (\mu\beta + \nu)\alpha^{N+1}, \\ A_n &= \nu(\beta^{N+1-n} - \alpha^{N+1-n}), \quad n=1, 2, \dots, N,\end{aligned}\quad (27)$$

For the no queue case, we obtain

$$\begin{aligned}p_0 &= \frac{(2\mu + \nu)(\mu + \nu) + \lambda\nu}{(2\mu + \nu)(\mu + \nu) + 2\lambda(\mu + \nu) + \lambda^2} \\ p_1 &= \frac{\lambda(2\mu + \nu)}{(2\mu + \nu)(\mu + \nu) + 2\lambda(\mu + \nu) + \lambda^2} \\ p_2 &= \frac{\lambda^2}{(2\mu + \nu)(\mu + \nu) + 2\lambda(\mu + \nu) + \lambda^2}\end{aligned}\quad (28)$$

And in the infinite queue case, under the condition $\lambda < 2(\mu + \nu)$, we have

$$\begin{aligned}\lim_{N \rightarrow \infty} p_0 &= \frac{\beta(\mu\alpha + \nu)[2(\mu + \nu) - \lambda]}{(\mu + \nu)[2\beta(\mu\alpha + \nu) + \nu]} \\ \lim_{N \rightarrow \infty} p_n &= \frac{\nu[2(\mu + \nu) - \lambda]}{\beta^{n-1}(\mu + \nu)[2\beta(\mu\alpha + \nu) + \nu]}, \quad n=1, 2, \dots,\end{aligned}\quad (29)$$

The mean number of customers in the system M/BVE/2(N) is

$$\begin{aligned}\sum_{n=0}^N np_n &= F'(1) = \{[(2\mu + \nu)(p_0 + p_1) - (N+2)\lambda p_N][2(\mu + \nu) - \lambda] \\ &\quad - [(\mu + \nu)(2p_0 + p_1) - \lambda p_N](2\mu + \nu - 2\lambda)\}[2(\mu + \nu) - \lambda]^{-2}.\end{aligned}\quad (30)$$

And the probability that both servers will be busy is

$$\sum_{n=2}^N p_n = \frac{\nu p_0}{(\mu\alpha + \nu)\beta^{N+1} - (\mu\beta + \nu)\alpha^{N+1}} \left(\frac{\beta^N - \beta^2}{\beta - 1} - \frac{\alpha^N - \alpha^2}{\alpha - 1} \right). \quad (31)$$

2.3. Optimal Assignment of service rates for Two Servers with General Input

Up to this section, we assumed that the service rate at which only one service finishes is all equal to μ . Here this rates are assumed to be μ_1 and μ_2 for two servers.

Namely, the service time distribution is the bivariate exponential distribution $BVE(\mu_1, \mu_2, \nu)$. The interarrival distribution will be arbitrary. Thus, we consider the queuing system $G/BVE(\mu_1, \mu_2, \nu)/2(N)$.

The problem is to seek an optimal $\mu_i (i = 1, 2)$, so as to minimize the loss call probability, under the condition $\mu_1 + \mu_2 = \mu(\text{constant})$.

2.3.1. Rate of Loss Calls

Let $A(t)$ be the interarrival distribution with arrival rate λ and $B_1(t), B_2(t), B_3(t)$ be the exponential distributions with parameters μ_1, μ_2 , and ν respectively.

We shall denote the states of this system as follows:

- $(i, j; x)$: number of customers in server I is i , in server II is $j (i, j = 0, 1)$ and the elapsed time from the last arrival is x .
- $(0; x)$: simple expression of $(0, 0; x)$
- $(n; x)$: the number of customers in the system is $n (n \geq 2)$ and the elapsed time from the last arrival is x .

Let $f(s; x)$ be the expected time until the first loss call, starting with the state $(s; x)$. Moreover, we shall use the following notations:

$$\begin{aligned} A_x(y) &= [A(y+x) - A(x)] / [1 - A(x)] \\ \overline{A}_x(y) &= 1 - A_x(y) \\ \overline{B}_i(y) &= 1 - B_i(y) \quad (i = 1, 2 \text{ and } 3) \end{aligned}$$

$B_i(y)AB_j(y)$ =distribution of minimum of $B_i(y)$ and $B_j(y)$ ($i, j=1, 2$ and $3, i \neq j$)

$$\overline{B_i(y)AB_j(y)} = 1 - (B_i(y)AB_j(y))$$

$B_1(y)AB_2(y)AB_3(y)$ =distribution of minimum of $B_1(y), B_2(y)$ and $B_3(y)$

$$\overline{B_1(y)AB_2(y)AB_3(y)} = 1 - (B_1(y)AB_2(y)AB_3(y)).$$

Using these notations, we can obtain the following equations in a straightforward manner.

$$\begin{aligned}
 f(0; x) &= \int_0^\infty [y + f(1, 0; 0)] dA_x(y) \\
 f(1, 0; x) &= \int_0^\infty [y + f(2, 0)] (\overline{B_1(y)AB_3(y)}) dA_x(y) \\
 &\quad + \int_0^\infty [y + f(0; x+y)] \overline{A_x(y)} d(B_1(y)AB_3(y)) \\
 f(0, 1; x) &= \int_0^\infty [y + f(2, 0)] (\overline{B_2(y)AB_3(y)}) dA_x(y) \\
 &\quad + \int_0^\infty [y + f(0; x+y)] \overline{A_x(y)} d(B_2(y)AB_3(y)) \\
 f(2; x) &= \int_0^\infty [y + f(3; 0)] (\overline{B_1(y)AB_2(y)AB_3(y)}) dA_x(y) \\
 &\quad + \int_0^\infty [y + f(0; x+y)] \overline{A_x(y)} (\overline{B_1(y)AB_2(y)}) dB_3(y) \\
 &\quad + \int_0^\infty [y + f(1, 0; x+y)] \overline{A_x(y)} (\overline{B_1(y)AB_3(y)}) dB_2(y) \\
 &\quad + \int_0^\infty [y + f(0, 1; x+y)] \overline{A_x(y)} (\overline{B_2(y)AB_3(y)}) dB_1(y) \\
 f(3; x) &= \int_0^\infty [y + f(4; 0)] (\overline{B_1(y)AB_2(y)AB_3(y)}) dA_x(y) \\
 &\quad + \int_0^\infty [y + f(1, 0; x+y)] \overline{A_x(y)} (\overline{B_1(y)AB_2(y)}) dB_3(y) \\
 &\quad + \int_0^\infty [y + f(2; x+y)] \overline{A_x(y)} \overline{B_3(y)} d(B_1(y)AB_2(y)) \\
 f(k; x) &= \int_0^\infty [y + f(k+1; 0)] (\overline{B_1(y)AB_2(y)AB_3(y)}) dA_x(y) \\
 &\quad + \int_0^\infty [y + f(k-2; x+y)] \overline{A_x(y)} (\overline{B_1(y)AB_2(y)}) dB_3(y) \\
 &\quad + \int_0^\infty [y + f(k-1; x+y)] \overline{A_x(y)} \overline{B_3(y)} d(B_1(y)AB_2(y)) \\
 &\quad (4 \leq k \leq N-1) \\
 f(N; x) &= \int_0^\infty y (\overline{B_1(y)AB_2(y)AB_3(y)}) dA_x(y) \\
 &\quad + \int_0^\infty [y + f(N-2; x+y)] \overline{A_x(y)} (\overline{B_1(y)AB_2(y)}) dB_3(y) \\
 &\quad + \int_0^\infty [y + f(N-1; x+y)] \overline{A_x(y)} \overline{B_3(y)} d(B_1(y)AB_2(y))
 \end{aligned} \tag{32}$$

These equations hold for $N \geq 4$. In case of $N = 2$, last three equations vanish and $f(3;x)$ in the equation of $f(2;x)$ must be zero. For $N = 3$, last two equations vanish and $f(4;x)$ in the equation of $f(3;x)$ must be zero.

For simplicity, we put

$$f(0;0)=f_0, f(1,0;0)=f_{10}, f(0,1;0)=f_{01}, f(n;0)=f_n \quad (n \geq 2) \quad (33)$$

Solving the system of equations (32), we obtain

$$\begin{aligned} f_0 &= \lambda^{-1} + f_{10} \\ a(\mu_1 + \nu)f_{10} &= \lambda^{-1} + a(\mu_1 + \nu)f_2 \\ f_{01} &= \lambda^{-1} + a(\mu_2 + \nu)f_2 + (1 - a(\mu_2 + \nu))f_{10} \\ f_m &= \lambda^{-1} + A_m a(\mu_1 + \nu) + B_m a(\mu_2 + \nu) \\ &+ \sum_{i=0}^{m-2} \left[\sum_{j=0}^{\min(i, m-i-2)} \mu^{m-i} \nu^j \binom{m}{i} f_{m-i-j+1} \right] (-1)^i a^{(i)}(\mu_1 + \mu_2 + \nu) \\ &- \sum_{i=0}^{m-2} \left[\sum_{j=0}^{\min(i, m-i-2)} \mu^{m-i} \nu^j \binom{m}{i} \right] (-1)^i a^{(i)}(\mu_1 + \mu_2 + \nu) f_{10} \\ &- \sum_{i=0}^{m-2} \left[\sum_{j=0}^{\min(i, m-i-2)} \mu^{m-i} \nu^j \binom{m}{i} (A_{m-i-j} + B_{m-i-j}) \right] (-1)^i a^{(i)}(\mu_1 + \mu_2 \\ &+ \nu) \end{aligned} \quad (34)$$

$$(2 \leq m \leq N; f_{N+1}=0)$$

where

$$\begin{aligned} a(s) &= \int_0^\infty \exp(-sx) dA_x(x) \\ a^{(i)}(s) &= \frac{d^i}{ds^i} a(s) \\ A_m &= a_1 \xi_1^m + a_2 \xi_2^m, B_m = b_1 \zeta_1^m + b_2 \zeta_2^m \end{aligned} \quad (35)$$

$$\begin{aligned} a_1 &= \frac{\xi_2 - \xi_1}{\lambda a(\mu_1 + \nu) \xi_1 (\xi_2 - \xi_1)}, \quad a_2 = \frac{1 - \xi_1}{\lambda a(\mu_1 + \nu) \xi_2 (\xi_2 - \xi_1)} \\ \xi_1 &= [\mu_1 + \mu_2 + \{(\mu_1 + \mu_2)^2 - 4\mu_2 \nu\}^{1/2}] / (2\mu_2) \\ \xi_2 &= [\mu_1 + \mu_2 - \{(\mu_1 + \mu_2)^2 - 4\mu_2 \nu\}^{1/2}] / (2\mu_2) \\ b_1 &= -[\lambda a(\mu_1 + \nu) \zeta_1 (\zeta_2 - \zeta_1)]^{-1}, \quad b_2 = [\lambda a(\mu_1 + \nu) \zeta_2 (\zeta_2 - \zeta_1)]^{-1} \\ \zeta_1 &= [(\mu_1 + \mu_2) + \{(\mu_1 + \mu_2)^2 - 4\mu_1 \nu\}^{1/2}] / (2\mu_1) \\ \zeta_2 &= [(\mu_1 + \mu_2) - \{(\mu_1 + \mu_2)^2 - 4\mu_1 \nu\}^{1/2}] / (2\mu_1). \end{aligned}$$

To find the rate of loss call, we have to solve the equations (34). If f_N is solved, then $1/(\lambda f_N)$ is the rate of loss call.

2.3.2. Optimal allocation for no queue case

In the system $G/BVE(\mu_1, \mu_2, \nu)/2(2)$, the probability of loss call becomes

$$P = \frac{\alpha(\mu_1 + \nu)\alpha(\mu + \nu)}{1 + \alpha(\mu + \nu) - \alpha(\mu - \mu_1 + \nu)} \quad (36)$$

In order to get the optimal value μ^* for μ_1 , we differentiate this equation with respect to μ_1 . Then we obtain μ^* as the unique root of $\beta(\mu^*) = 0$, only when $\beta(\mu)$ is positive, where

$$\beta(\mu_1) = \alpha^{(1)}(\mu_1 + \nu)[1 + \alpha(\mu + \nu) - \alpha(\mu - \mu_1 + \nu)] - \alpha(\mu_1 + \nu)\alpha^{(1)}(\mu - \mu_1 + \nu). \quad (37)$$

This μ^* exists between $\mu/2$ and μ . When $\beta(\mu)$ is negative, optimal value of μ_1 is μ (μ_2 is 0).

Now, we shall show the optimal allocations for some type of arrival distributions.

(a) Random Arrivals. ($M/MVE(\mu_1, \mu_2, \nu)/2(2)$)

When

$$0 < \lambda < [-(\mu - \nu) + \{(\mu + \nu)^2 + 4\lambda\nu\}^{1/2}]/2$$

the optimal service rates are

$$\mu_1 = \mu, \quad \mu_2 = 0.$$

And if

$$\lambda > [-(\mu - \nu) + \{(\mu + \nu)^2 + 4\lambda\nu\}^{1/2}]/2$$

the optimal service rates are

$$\mu_1 = \mu^*, \quad \mu_2 = \mu - \mu^*$$

where

$$\mu^* = (\lambda + \mu + \nu) - \left\{ \frac{(\lambda + \mu + \nu)(2\lambda + 2\nu + \mu)}{2\lambda + \mu + \nu} \right\}^{1/2} \quad (38)$$

(b) Erlangian Arrivals ($E_k/MVE(\mu_1, \mu_2, \nu)/2(2)$)

In this case, we have

$$\begin{aligned} \beta(\mu) = & -\frac{1}{(\mu + \nu + l\lambda)^{l+1}} \left\{ 1 - \left(\frac{l\lambda}{\nu + l\lambda} \right)^l \right\} \\ & + \left(\frac{l\lambda}{\mu + \nu + l\lambda} \right)^l \left\{ -\frac{1}{(\mu + \nu + l\lambda)^{l+1}} + \frac{1}{(\mu + l\lambda)^{l+1}} \right\} \end{aligned}$$

If $\beta(\mu) > 0$,

$$\mu^* = \mu + \nu + l\lambda - \left\{ \frac{(\mu + \nu + l\lambda)^l (l\lambda)^l (\mu + 2\nu + 2l\lambda)}{(\mu + \nu + l\lambda)^l + (l\lambda)^l} \right\}^{1/(l+1)} \quad (39)$$

is the optimal value of μ_1 , and if $\beta(\mu) < 0$, the optimal rates are $\mu_1 = \mu$, $\mu_2 = 0$.

(c) Regular Arrivals (D/MVE(μ_1 , μ_2 , ν)/2(2))

When

$$2 > \exp(-\mu/\lambda) + \exp(-\nu/\lambda)$$

the optimal service rates are $\mu_1 = \mu^*$, $\mu_2 = \mu - \mu^*$, where

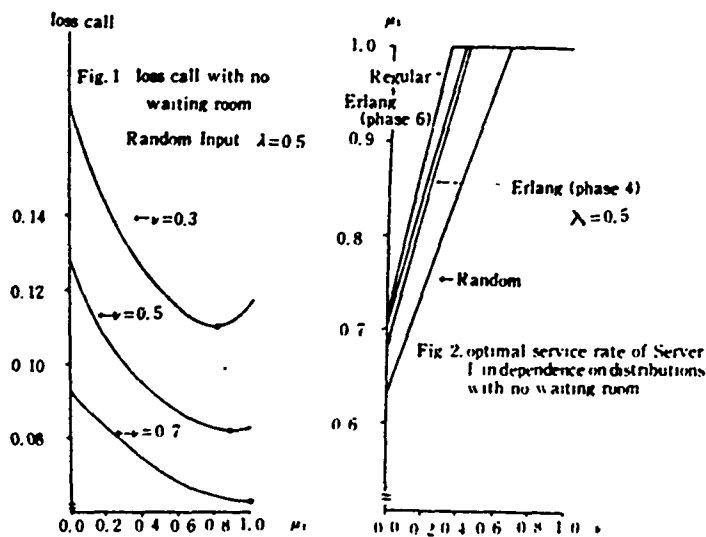
$$\mu^* = -\lambda \log 2 + \mu + 2\nu + \lambda \log [\exp(-\nu/\lambda) + \exp(-(\mu + 2\nu)/\lambda)]. \quad (40)$$

And when

$$2 < \exp(-\mu/\lambda) + \exp(-\nu/\lambda)$$

$\mu_1 = \mu$, $\mu_2 = 0$ are the optimal rates.

Fig. 2. G/MVE(μ_1 , μ_2 , ν)/2(2)



$$\mu_1 + \mu_2 = 1$$

2.3.3. Optimal Allocation with Random Input (M/MVE(μ_1, μ_2, ν)/2(N))

The loss call probability becomes

$$P = \lambda(\xi - \zeta)(\mu + 2\nu - \lambda)/E \quad (41)$$

where

$$\begin{aligned} E = & -(\mu_1 - \mu_2)(\mu + 2\nu - \lambda)\mu_1(\xi^{N-1} - \zeta^{N-1}) \\ & -(\lambda + \mu_2 + \nu)[\lambda^2(\xi - \zeta) + \{\mu_2(\mu + 2\nu) - \lambda(\mu_2 + \nu)\}(\xi^N - \zeta^N) \\ & -(\mu + 2\nu)\lambda(\xi^{N+1} - \zeta^{N+1})]. \end{aligned}$$

and

$$\begin{aligned} \xi &= [\mu_1 + \mu_2 + \nu - \{(\mu_1 + \mu_2 + \nu)^2 + 4\lambda\nu\}^{1/2}]/(2\lambda) \\ \zeta &= [\mu_1 + \mu_2 + \nu + \{(\mu_1 + \mu_2 + \nu)^2 + 4\lambda\nu\}^{1/2}]/(2\lambda) \end{aligned}$$

After some simple calculations, we have the following results. Let

$$\sigma(x) = x^2 - 2(\lambda + \mu + \nu)x + \frac{(\lambda + \mu + \nu)\{\mu(\xi^{N-1} - \zeta^{N-1}) + (\lambda + \mu + \nu)(\xi^N - \zeta^N)\}}{2(\xi^{N-1} - \zeta^{N-1}) + (\xi^N - \zeta^N)} \quad (42)$$

then if $\sigma(\mu) > 0$, the optimal service rates are $\mu_1 = \mu$, $\mu_2 = 0$. And if $\sigma(\mu) < 0$, $\mu_1 = \mu^*$, $\mu_2 = \mu - \mu^*$ are optimal, where

$$\mu^* = (\lambda + \mu + \nu) - \left\{ \frac{(2\lambda + \mu + \nu)(\xi^{N-1} - \zeta^{N-1})(\lambda + \mu + \nu)}{2(\xi^{N-1} - \zeta^{N-1}) + (\xi^N - \zeta^N)} \right\}^{1/2} \quad (43)$$

Letting $\nu \rightarrow 0$, we can get the optimal allocation for M/M/2(N) and the optimal value of μ_1 is

$$\lim_{\nu \rightarrow 0} \mu^* = \lambda + \mu - \left\{ \frac{(2\lambda + \mu)\zeta^{N-1}(\lambda + \mu)}{2\zeta^{N-1} + \zeta^N} \right\}^{1/2} = \lambda + \mu - \{(\lambda + \mu)\lambda\}^{1/2} \quad (44)$$

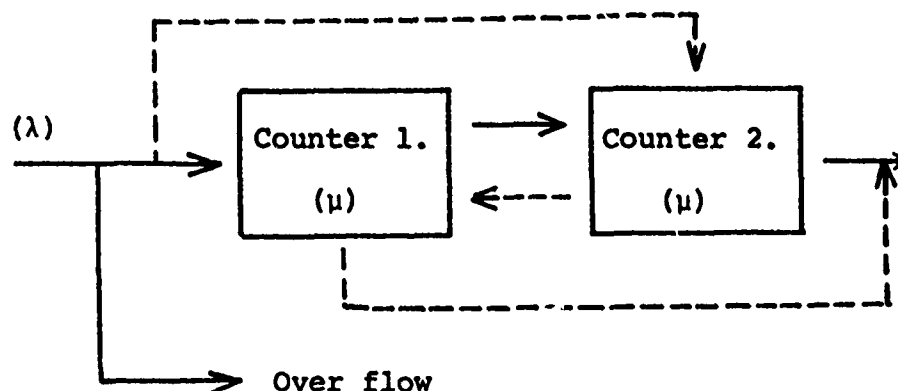
3. COMMUTATIVE TANDEM QUEUE

In some assembly lines, there are many cases in which empty stations can be used regardless of ordered sequence of stations in order to increase efficiency of system. We shall call such system commutative tandem queue. We consider a two serial stations. We assume that customers arrive according to a Poisson stream with parameter λ and each service time of two stations is exponentially distributed with same parameter μ . In the case of which each

service rate of two stations is different, the detailed balance equations for steady states are easily derived but its analysis is complicated. So, for simplicity, we only concern with the case of the same service rate.

Moreover, we assume that there are no queue between two stations.

Fig. 3. Commutative tandem queue



Arriving customers enter the first station if both stations are free, and they join the queue if both are occupied. They can first enter the second station if this is free and the first station is busy. If a customer has already completed service of two stations, then he emerges from this system. But if he has not completed by the other station and it is not free, he has to stay there, that is to say, this station is blocked and the other station has completed service, he is able to enter it.

It is also assumed that customers can transfer between stations instantaneously. The queuing discipline is first-come-first-served.

3.1. Some Characteristics for Finite Waiting Room

We shall consider a commutative tandem queue in which the waiting room allowed ahead of the first station is finite. The capacity of waiting room is N . A customer who, upon his arrival, finds the system full departs never to return.

The particular state of the system is labeled by the states of the queue length ahead of the first station and the states of the two stations. The state of the queue length is represented by the number of customers in queue. Each station can be empty (0), serving a customer who has

not received (unfinished) service at the other station (u), serving a customer who finished service already at the other station (f), or blocked (b) when it has completed own service but the other is still occupied. It is convenient to express the probability for this system by the form $P(\cdot, \cdot, \cdot)$, where the first dot is the state of queue, the second is that of the first station and the third is that of the second station. For simplicity, the probability of no customers in the system is denoted by $P(0)$.

The detailed balance equations for steady-states are as follows.

$$\begin{aligned}
 \lambda P(0) &= \mu P(0, f, 0) + \mu P(0, 0, f) \\
 (\lambda + \mu) P(0, u, 0) &= \lambda P(0) + \mu P(0, u, f) \\
 (\lambda + \mu) P(0, f, 0) &= \mu P(0, 0, u) + \mu P(0, f, f) + \mu P(0, f, b) \\
 (\lambda + \mu) P(0, 0, u) &= \mu P(0, f, u) \\
 (\lambda + \mu) P(0, 0, f) &= \mu P(0, u, 0) + \mu P(0, f, f) + \mu P(0, b, f) \\
 (\lambda + 2\mu) P(0, u, u) &= \lambda P(0, 0, u) + \lambda P(0, u, 0) + \mu P(1, f, u) + \mu P(1, u, f) \\
 (\lambda + 2\mu) P(0, f, f) &= \mu P(0, u, b) + \mu P(0, b, u) \\
 (\lambda + 2\mu) P(0, u, f) &= \lambda P(0, 0, f) + \mu P(1, f, f) + \mu P(1, b, f) \\
 (\lambda + 2\mu) P(0, f, u) &= \lambda P(0, f, 0) + \mu P(1, f, f) + \mu P(1, f, b) \\
 (\lambda + \mu) P(0, b, u) &= \mu P(0, u, u) \\
 (\lambda + \mu) P(0, b, f) &= \mu P(0, u, f) \\
 (\lambda + \mu) P(0, u, b) &= \mu P(0, u, u) \\
 (\lambda + \mu) P(0, f, b) &= \mu P(0, f, u)
 \end{aligned}
 \tag{45}$$

$$\begin{aligned}
 (\lambda + 2\mu) P(n, u, u) &= \lambda P(n-1, u, u) + \mu P(n+1, f, u) + \mu P(n+1, u, f) \\
 (\lambda + 2\mu) P(n, f, f) &= \lambda P(n-1, f, f) + \mu P(n, b, u) + \mu P(n, u, b) \\
 (\lambda + 2\mu) P(n, u, f) &= \lambda P(n-1, u, f) + \mu P(n+1, f, f) + \mu P(n+1, b, f) \\
 (\lambda + 2\mu) P(n, f, u) &= \lambda P(n-1, f, u) + \mu P(n+1, f, f) + \mu P(n+1, f, b) \\
 (\lambda + \mu) P(n, b, u) &= \lambda P(n-1, b, u) + \mu P(n, u, u) \\
 (\lambda + \mu) P(n, b, f) &= \lambda P(n-1, b, f) + \mu P(n, u, f) \\
 (\lambda + \mu) P(n, u, b) &= \lambda P(n-1, u, b) + \mu P(n, u, u) \\
 (\lambda + \mu) P(n, f, b) &= \lambda P(n-1, f, b) + \mu P(n, f, u)
 \end{aligned}
 \tag{1 \leq n \leq N-1}$$

$$\begin{aligned}
 2\mu P(N, u, u) &= \lambda P(N-1, u, u) \\
 2\mu P(N, f, f) &= \lambda P(N-1, f, f) + \mu P(N, b, u) + \mu P(N, u, b) \\
 2\mu P(N, u, f) &= \lambda P(N-1, u, f) \\
 2\mu P(N, f, u) &= \lambda P(N-1, f, u) \\
 \mu P(N, b, u) &= \lambda P(N-1, b, u) + \mu P(N, u, u) \\
 \mu P(N, b, f) &= \lambda P(N-1, b, f) + \mu P(N, u, f) \\
 \mu P(N, u, b) &= \lambda P(N-1, u, b) + \mu P(N, u, u) \\
 \mu P(N, f, b) &= \lambda P(N-1, f, b) + \mu P(N, f, u)
 \end{aligned}$$

Setting

$$\begin{aligned}
 P(n,1) &= P(n,u,u) \\
 P(n,2) &= P(n,f,u) + P(n,u,f) \\
 P(n,3) &= P(n,f,f) \\
 P(n,4) &= P(n,u,b) + P(n,b,u) \\
 P(n,5) &= P(n,f,b) + P(n,b,f) ,
 \end{aligned} \tag{46}$$

we get

$$\begin{aligned}
 (2+p)P(n,1) &= \rho P(n-1,1) + P(n+1,2) \\
 (2+p)P(n,2) &= \rho P(n-1,2) + 2P(n+1,3) + P(n+1,5) \\
 (2+p)P(n,3) &= \rho P(n-1,3) + P(n,4) \\
 (1+p)P(n,4) &= \rho P(n-1,4) + 2P(n,1) \\
 (1+p)P(n,5) &= \rho P(n-1,5) + P(n,2) \quad (0 \leq n \leq N-1) \\
 2P(N,1) &= \rho P(N-1,1) \\
 2P(N,2) &= \rho P(N-1,2) \\
 2P(N,3) &= \rho P(N-1,3) + P(N,4) \\
 P(N,4) &= \rho P(N-1,4) + 2P(N,1) \\
 P(N,5) &= \rho P(N-1,5) + P(N,2) ,
 \end{aligned} \tag{47}$$

where

$$\begin{aligned}
 \rho &= \lambda/\mu \\
 P(-1,1) &= (\rho P(0) + P(0,2))/(\rho+1) \\
 P(-1,2) &= \rho P(0)
 \end{aligned}$$

$$P(-1,3) = P(-1,4) = P(-1,5) = 0$$

Now we define the five generating functions

$$G_i(z) = \sum_{n=0}^N z^n P(n,i) \quad (i=1,2,\dots,5) . \tag{48}$$

From the equations (47), we have

$$\begin{aligned}
 (2+p-\rho z)G_1(z) - G_4(z) &= \rho(1-z)z^N P(N,3) \\
 2G_1(z) - (1+p-\rho z)G_4(z) &= -\rho(1-z)z^N P(N,4) \\
 G_2(z) - (1+p-\rho z)G_5(z) &= -\rho(1-z)z^N P(N,5)
 \end{aligned} \tag{49}$$

The normalization equations is

$$\sum_{i=1}^5 G_i(1) + \sum_{j=1}^5 P(-1,j) + P(0) = 1 . \tag{50}$$

We can solve $G_i(z)$ ($i = 1, \dots, 5$) using (49), (50) and the regularity of $G_i(z)$. And we obtain

$$G_i(z) = \frac{H_i(z)}{F(z)} \quad (i=1,2), \quad G_i(z) = \frac{H_i(z)}{(1+p-\rho z)F(z)} \quad (i=3,4,5), \tag{51}$$

where

$$F(z) = (\rho+1)(-\rho^4 z^5 + (3\rho^4 + 7\rho^3)z^4 - (3\rho^4 + 14\rho^3 + 18\rho^2)z^3 + (\rho^4 + 7\rho^3 + 19\rho^2 + 20\rho)z^2 - (\rho^2 + 4\rho + 8)z - 4).$$

$H_i(z)$ ($i = 1, \dots, 5$) can be also expressed explicitly, but these expressions are lengthy. So we shall omit here these one.

The mean length in the queue is

$$\begin{aligned} L_c &= G_1'(1) + G_2'(1) + G_3'(1) + G_4'(1) + G_5'(1) \\ &= \frac{7\rho + 16N + 24}{16(4\rho - 3)(\rho + 1)}P(0, 2) + \frac{19\rho^3 + (16N + 47)\rho^2 + (80N + 181)\rho + (48N + 136)}{16(4\rho - 3)(\rho + 1)}P(0) \\ &\quad + \frac{128\rho^2 - (64N + 227)\rho + (48N + 136)}{16(3 - 4\rho)} + 2\rho P(N, 1) + \frac{8\rho(1 - \rho)}{(3 - 4\rho)}P(N, 2) \\ &\quad + \frac{21\rho - 16\rho^2}{6 - 8\rho}P(N, 3) + \frac{5\rho - 16\rho^2}{6 - 8\rho}P(N, 4) + \frac{9\rho - 8\rho^2}{3 - 4\rho}P(N, 5), \end{aligned} \quad (52)$$

and the mean availability per station is

$$\begin{aligned} A_c &= \frac{1}{2} \left(2 \sum_{i=1}^3 G_i(1) + \sum_{i=4}^5 G_i(1) + \sum_{j=1}^2 P(-1, j) \right) \\ &= \frac{3}{4} - \frac{P(0, 2)}{4(\rho + 1)} - \frac{\rho^2 + 5\rho + 3}{4(\rho + 1)} \end{aligned} \quad (53)$$

The blocking probability is

$$P_{BC} = \frac{1}{2} - \frac{P(0, 2)}{2(\rho + 1)} - \frac{\rho^2 + 3\rho + 1}{2(\rho + 1)} \quad (54)$$

3.2. Infinite Queue Case

Here we shall admit infinite possible queue ahead of the first station. Other conditions are the same as preceding section.

In this case, we have

$$G_i(z) = \frac{H_i(z)}{F(z)} \quad (i = 1, 2, \dots, 5) \quad (55)$$

where

$$F(z) = (\rho + 1) \{ -\rho^4 z^5 + (3\rho^4 + 7\rho^3)z^4 - (3\rho^4 + 14\rho^3 + 18\rho^2)z^3 + (\rho^4 + 7\rho^3 + 19\rho^2 + 20\rho)z^2 - (\rho^2 + 4\rho + 8)z - 4 \},$$

$$\begin{aligned}
H_1(z) &= (2 + \rho - \rho z) [[-\rho^4 z^3 + (\rho^4 + 3\rho^3)z^2 + (\rho^4 + \rho^3 - 2\rho^2)z - (\rho^4 + 3\rho^3 + 2\rho^2)] P(0) \\
&\quad + [-\rho^3 z^3 + (2\rho^3 + 4\rho^2)z^2 - (\rho^3 + 4\rho^2 + 5\rho)z + (2\rho + 2)] P(0, 2)], \\
H_2(z) &= [\rho^5(\rho + 1)z^4 - (3\rho^6 + 9\rho^5 + 5\rho^4)z^3 + (3\rho^6 + 15\rho^5 + 23\rho^4 + 8\rho^3)z^2 \\
&\quad - (\rho^6 + 7\rho^5 + 18\rho^4 + 20\rho^3 + 4\rho^2)z] P(0) + [\rho^3 z^3 - (2\rho^3 + 5\rho^2)z^2 \\
&\quad + (\rho^3 + 5\rho^2 + 8\rho)z - 4\rho - 4] P(0, 2),
\end{aligned}$$

$$\begin{aligned}
H_3(z) &= \frac{2}{1 + \rho - \rho z} [[-\rho^4 z^3 + (\rho^4 + 3\rho^3)z^2 + (\rho^4 + \rho^3 - 2\rho^2)z - (\rho^4 + 3\rho^3 + 2\rho^2)] P(0) \\
&\quad + [-\rho^3 z^3 + (2\rho^3 + 4\rho^2)z^2 - (\rho^3 + 4\rho^2 + 5\rho)z + (2\rho + 2)] P(0, 2)],
\end{aligned}$$

$$\begin{aligned}
H_4(z) &= \frac{2(2 + \rho - \rho z)}{1 + \rho - \rho z} [[-\rho^4 z^3 + (\rho^4 + 3\rho^3)z^2 + (\rho^4 + \rho^3 - 2\rho^2)z - (\rho^4 + 3\rho^3 + 2\rho^2)] P(0) \\
&\quad + [-\rho^3 z^3 + (2\rho^3 + 4\rho^2)z^2 - (\rho^3 + 4\rho^2 + 5\rho)z + (2\rho + 2)] P(0, 2)],
\end{aligned}$$

$$\begin{aligned}
H_5(z) &= \frac{1}{1 + \rho - \rho z} [[\rho^5(\rho + 1)z^4 - (3\rho^6 + 9\rho^5 + 5\rho^4)z^3 + (3\rho^6 + 15\rho^5 + 23\rho^4 + 8\rho^3)z^2 \\
&\quad - (\rho^6 + 7\rho^5 + 18\rho^4 + 20\rho^3 + 4\rho^2)z] P(0) + [\rho^3 z^3 - (2\rho^3 + 5\rho^2)z^2 + (\rho^3 + 5\rho^2 + 8\rho)z \\
&\quad - 4\rho - 4] P(0, 2)].
\end{aligned}$$

The mean queue length is

$$\begin{aligned}
L_c &= G'_1(1) + G'_2(1) + G'_3(1) + G'_4(1) + G'_5(1) \\
&= \frac{3\rho^2 - 6\rho + 16}{4(4\rho - 3)} P(0) + \frac{-39\rho^2 - 38\rho + 16}{4(4\rho - 3)} \quad (0 \leq \rho < \frac{3}{4}), \quad (56)
\end{aligned}$$

and the mean availability per station is

$$\begin{aligned}
A_c &= \frac{1}{2} [2(G_1(1) + G_2(1) + G_3(1)) + G_4(1) + G_5(1) + P(-1, 1) + P(1, 2)] \\
&= \rho \quad (0 \leq \rho < \frac{3}{4}). \quad (57)
\end{aligned}$$

The blocking probability is

$$P_B = P(0) + 2\rho - 1 \quad (0 \leq \rho < \frac{3}{4}) \quad (58)$$

Now, we shall display the mean queue length in order to compare commutative with ordinary tandem queue for $N = 0, 1, 2, \infty$. Here L_0 is the mean queue length in the ordinary

tandem queuing system.

TABLE 2

p \ N	1		2		∞	
	L_C	L_O	L_C	L_O	L_C	L_O
0.1	0.003	0.010	0.004	0.013	0.004	0.013
0.2	0.018	0.040	0.027	0.061	0.031	0.073
0.3	0.048	0.084	0.082	0.151	0.116	0.228
0.4	0.090	0.136	0.172	0.278	0.324	0.600
0.5	0.139	0.192	0.290	0.426	0.821	1.600
0.6	0.191	0.248	0.424	0.582	2.208	6.092
0.7	0.242	0.301	0.563	0.731	9.878	

From this table, it is seen that the mean queue length of commutative system is smaller than the ordinary one for each ρ . Thus, the efficiency of commutative system is better than the ordinary's.

3.3. Commutative Tandem Queue with Correlated Two Servers

We assume that the service times of two servers are the bivariate exponential distribution $BVE(\mu, \mu, \nu)$, and unlimited queue is allowed ahead of the first station and no queue is allowed between two stations.

The detailed balance equations for steady states are as follows:

$$\begin{aligned}
 \lambda P(0) &= (\mu + \nu)P(0, f, 0) + (\mu + \nu)P(0, 0, f) + \nu P(0, f, f) \\
 (\lambda + \mu + \nu)P(0, u, 0) &= \lambda P(0) + \mu P(0, u, f) + \nu P(1, f, f) \\
 (\lambda + \mu + \nu)P(0, f, 0) &= (\mu + \nu)P(0, 0, u) + \mu P(0, f, f) \\
 &\quad + (\mu + \nu)P(0, f, b) + \nu P(0, f, u) \\
 (\lambda + \mu + \nu)P(0, 0, u) &= \mu P(0, f, u) \\
 (\lambda + \mu + \nu)P(0, 0, f) &= (\mu + \nu)P(0, u, 0) + \mu P(0, f, f) \\
 &\quad + (\mu + \nu)P(0, b, f) + \nu P(0, u, f) \\
 (\lambda + 2\mu + \nu)P(0, u, u) &= \lambda P(0, 0, u) + \lambda P(0, u, 0) + \mu P(1, f, u) \\
 &\quad + \mu P(1, u, f) + \nu P(2, f, f)
 \end{aligned}$$

$$\begin{aligned}
(\lambda+2\mu+\nu)P(0,f,f) &= (\mu+\nu)P(0,u,b) \\
&\quad + (\mu+\nu)P(0,b,u) + \nu P(0,u,u) \\
(\lambda+2\mu+\nu)P(0,u,f) &= \lambda P(0,0,f) + \mu P(1,f,f) \\
&\quad + (\mu+\nu)P(1,b,f) + \nu P(1,u,f) \quad (59) \\
(\lambda+2\mu+\nu)P(0,f,u) &= \lambda P(0,f,0) + \mu P(1,f,f) \\
&\quad + (\mu+\nu)P(1,f,b) + \nu P(1,f,u) \\
(\lambda+\mu+\nu)P(0,b,u) &= \mu P(0,u,u) \\
(\lambda+\mu+\nu)P(0,b,f) &= \mu P(0,u,f) \\
(\lambda+\mu+\nu)P(0,u,b) &= \mu P(0,u,u) \\
(\lambda+\mu+\nu)P(0,f,b) &= \mu P(0,f,u)
\end{aligned}$$

By the same manner as preceding sections, we have

$$\begin{aligned}
\theta G_1(z) - (2+\rho+\theta-\rho z)G_3(z) + (1+\theta)G_4(z) &= 0 \\
2G_1(z) - (1+\rho+\theta-\rho z)G_4(z) &= 0 \\
G_2(z) - (1+\rho+\theta-\rho z)G_5(z) &= 0
\end{aligned} \quad (60)$$

where

$$\rho = \lambda/\mu, \quad \theta = \nu/\mu.$$

We can solve $G_i(z)$ ($i = 1, \dots, 5$) also similarly as preceding sections.

The mean queue length is

$$\begin{aligned}
L_c &= \sum_{i=1}^5 G_i'(1) \\
&= (4 + \frac{4}{1+\theta})G_1'(1) \\
&\quad + \frac{(2\theta^4 + 12\theta^3 + 26\theta^2 + 24\theta - 2\rho\theta^3 - 9\rho\theta^2 - 6\rho\theta + 7\rho + 8)}{(2+\theta)(1+\theta)^2} G_1(1) \\
&\quad + \frac{(-\rho^3 - \rho^3\theta - 3\rho^2\theta - \rho^2\theta^2 - 2\rho^2)}{(1+\theta)^3} P(0) + \left(\frac{\rho\theta + \theta + 1}{1+\theta} \frac{1+\theta}{\rho+1+\theta} \right) P(0,2)
\end{aligned}$$

$$\begin{aligned}
& + (-4 - 2\theta - \frac{\rho\theta^3 + 3\rho\theta^2 + 4\rho\theta + 2\rho + \rho^2\theta}{(1+\theta)^3}) \\
& + \frac{(\rho+1+\theta)(\rho\theta+2+3\theta+\theta^2)(2+\theta)}{(1+\theta)^3} P(0,3) , \\
& (0 \leq \rho < \frac{(1+\theta)(2\theta+3)}{2\theta+4}) .
\end{aligned} \tag{61}$$

and the mean availability per station is

$$\begin{aligned}
A_c &= \frac{1}{2} (2 \sum_{i=1}^3 G_i(1) + \sum_{i=4}^5 G_i(1) + \sum_{j=1}^2 P(-1, j)) \\
&= \frac{1}{2} ((8 + \frac{4}{1+\theta}) G_1(1) \\
&+ ((2\rho+1 + \frac{\rho}{1+\theta}) \frac{\rho}{1+\theta} + \frac{\rho}{\rho+1+\theta} + \frac{\rho^2(\rho+2+2\theta)}{(\rho+1+\theta)(1+\theta)^2}) P(0) \\
&- ((2 + \frac{1}{1+\theta})(\theta + \frac{1+\theta}{\rho+1+\theta}) + \frac{\theta^2 + \rho\theta + 2\theta + 1}{(\rho+1+\theta)(1+\theta)}) P(0,2) \\
&- (\frac{6\theta^2 + 14\theta + 2\rho\theta^2 + 4\rho\theta + 8}{(1+\theta)^2}) P(0,3) , \\
&(0 \leq \rho < \frac{(1+\theta)(2\theta+3)}{2\theta+4}) .
\end{aligned} \tag{62}$$

In this system the maximum utilization factor is

$$\rho_{\max} = \frac{(1+\theta)(2\theta+3)}{(2\theta+4)} \tag{63}$$

TABLE 3

v	$\lambda=0.1, \quad \mu=1.0$				$\lambda=0.3, \quad \mu=1.0$			
	L_C	L_O	A_C	A_O	L_C	L_O	A_C	A_O
1.0	.0005	.0028	.0281	.0498	.0068	.0320	.1146	.1483
0.9	.0006	.0031	.0309	.0524	.0084	.0364	.1215	.1560
0.8	.0007	.0035	.0341	.0553	.0105	.0417	.1287	.1645
0.7	.0008	.0040	.0377	.0585	.0133	.0483	.1368	.1740
0.6	.0010	.0046	.0419	.0622	.0171	.0566	.1464	.1847
0.5	.0012	.0053	.0470	.0663	.0222	.0672	.1583	.1969
0.4	.0015	.0062	.0532	.0710	.0295	.0812	.1736	.2109
0.3	.0018	.0073	.0608	.0765	.0403	.1002	.1937	.2272
0.2	.0023	.0087	.0706	.0829	.0571	.1268	.2211	.2466
0.1	.0029	.0107	.0835	.0906	.0844	.1661	.2594	.2702
0.0	.0037	.0135	.1000	.1000	.1161	.2277	.3000	.3000

REFERENCES

- [1] Marshall, A.W. and I. Olkin, MULTIVARIATE EXPONENTIAL DISTRIBUTION, J. Am. Stat. Assoc., Vol 62, PP. 30-44, 1967
- [2] Nishida, T., A. Tahara and H. Hanai, OPTIMAL DESIGN FOR HETEROGENOUS TWO-SERVER QUEUE, Tech. Rept. Osaka Univ., Vol 22, PP. 295-301, 1972
- [3] Nishida, T., R. Watanabe and A. Tahara, POISSON QUEUE WITH CORRELATED TWO SERVERS, Tech. Rept. Osaka Univ., Vol 24, PP. 403-409, 1974
- [4] Nishida, T. and H. Kubota, OPTIMAL ALLOCATION OF SERVICE RATES FOR CORRELATED TWO-SERVER QUEUE WITH GENERAL INPUT, Math. Japonicae, Vol 20, PP. 161-170, 1975
- [5] Nishida, T., T. Hiramatsu and H. Itsuji, COMMUTATIVE TANDEM QUEUE WITH MARKOVIAN INPUT AND SERVICES, Math. Japonicae, Vol 21, PP. 401-409, 1976
- [6] Nishida, T. and K. Yoneyama, CORRELATED MULTISERVER QUEUE WITH RANDOM INPUT, Math. Japonicae, Vol 22, PP.395-401, 1977
- [7] Nishida, T. and T. Hiramatsu, COMMUTATIVE TANDEM QUEUE WITH FINITE WAITING ROOM, J. Oper. Res. Soc. Japan, Vol 20, PP. 194-202, 1977
- [8] Nishida, T. and T. Hiramatsu, COMMUTATIVE TANDEM QUEUE WITH CORRELATED TWO SERVERS, to appear in Math. Japonicae

**Methodological and Modelling Approaches
for Projecting Health Manpower
Requirements and Supply**

by

**Kong-Kyun Ro
Korea Advanced Institute of Science**

**An Invited Paper
Delivered**

at

**The Pacific Conference on Operations Research
April 23-28, 1979
Seoul, Korea**

Sponsored by

**Military Operations Research Society of Korea
and
Korean Operations Research Society**

INTRODUCTION

Will Korea have a sufficient number of engineers in 1985? Does Korea have a shortage of doctors today and will it become more acute in 1980's? These are important questions to be raised in manpower planning. Various methods are available to make projections of the future manpower supply and requirements to answer these questions. No one method is, however, universally accepted and free of errors. Experiences has shown that different methods produce different projections. For example, in the U.S.A. where the art or science of manpower projection is supposed to be relatively more perfected, six projections of physician requirements based on different methods produced the estimates which varied from 305,000 to 425,000.^{1/} Therefore, these findings led to opposing conclusions concerning the adequacy of the projected supply of physicians in the U.S.A.

It is clear then that a manpower planner must be knowledgeable about the various methods available to project the future manpower supply and requirements so that he or she may choose an appropriate method to make projections or make an appropriate interpretation of the projections made by others using different methods. The purpose of this paper is to provide the needed information by discussing the state of art in the manpower projection methodologies as applied to the health field in the layman's language. The plan of the paper is to discuss the most commonly used methods in estimating and projecting health manpower requirements and supply ^{2/}and evaluate the relative strengths and weaknesses of each

^{1/} Lee W. Hansen, "An Appraisal of Physician Manpower Projections," Inquiry, March 1970.

^{2/} The typology of methodological approaches was adopted from Thomas L. Hall, "Estimating Requirements and Supply: Where Do We Stand?" (in the Pan American Conference on Health Planning, 10-17 September 1973, Washington, D.C.; Pan American Organization, Scientific Publication No. 279, 1974.

approach. The discussion and evaluation will be done with examples of how these methods are used in the four health manpower models which a former colleague of mine and I evaluated for the U.S. Department of Health, Education and Welfare under a contract.

MANPOWER-POPULATION RATIO METHOD

The methods most commonly used in estimating manpower requirements in the health field are those based on: (1) the manpower-population ratio; (2) the amount of service targeted to be provided; (3) the need of population; and (4) the effective demand. Manpower to population ratio approach is most frequently used in the health field perhaps because of its simplicity. This ratio approach selects a ratio of the number of health personnel to the total population. Then, the manpower requirements are calculated by applying the ratio to the target year population as follows:

$$\frac{\text{manpower}}{\text{population}} \times \text{target population} = \text{estimated manpower requirements.}$$

Given the above basic formula, it can be used to fit the particular circumstances of the user who is likely to be a manpower planner. For example, the numerator in the ratio, manpower, may represent an occupation or a sub-category of occupation or an all-inclusive profession. The denominator of the ratio, population, may represent a particular group of people with certain demographic characteristics or total community population.

The simplicity of the formula is the strong point of this ratio method. This simplicity, however, brings with it its own pitfalls. This is because the validity of this method depends entirely on how the ratio is chosen and how it is used. A rigid application of this method is obviously likely to bring an inaccurate and unreliable estimate of manpower requirements. The population size is an important factor which explains manpower requirements of a given area. But the size itself is one of many factors which determine manpower requirements. Dif-

ferences in population characteristics, income and educational level of the populations and the productivity of manpower bring about the differences in manpower requirements between two areas of equal population size. The user of this model, therefore, must take account of how these other variables operate in his situation. In spite of these shortcomings, more sophisticated methodologies use this manpower-population ratio as an input or a principal input. One of the four models examined in this paper, namely, the Vector model ¹, uses the ratio approach as a major input.

Vector model : The Vector model uses demographic characteristics, income level and others as the "modifying" factors in projecting the nursing manpower requirements based on the manpower-population ratio. The model uses these variables to account for the differences in the per capita demand for health services. Therefore, it can be said that the model uses the demand approach as well as the manpower-population ratio method. In order to illustrate how a modelling approach to manpower requirements may use the manpower-population ratio method, the Vector model is described as an example in some details.

As shown in Figure 1, the Vector model consists of three modules: a population module, a demand-for-services module, and a nurse manpower requirements module. The demand-for-services module uses the projections from the population module to produce projections of health service demands by provider setting. The nurse manpower requirements module uses these projected health service demands to estimate future nursing requirements by employment setting.

The population module uses as input four types of data: (1) projections of the future U.S. population in terms of age, sex, and family status; (2) income distribution of the population; (3) health insurance coverage fractions for each population cohort; and (4) an Health Maintenance Organization (HMO) formation rate. These data are used by the

1. The Impact of Health Care System Changes on the Nation's Requirements for Registered Nurses in 1985, Vector Research, Inc., Ann Arbor, Michigan.

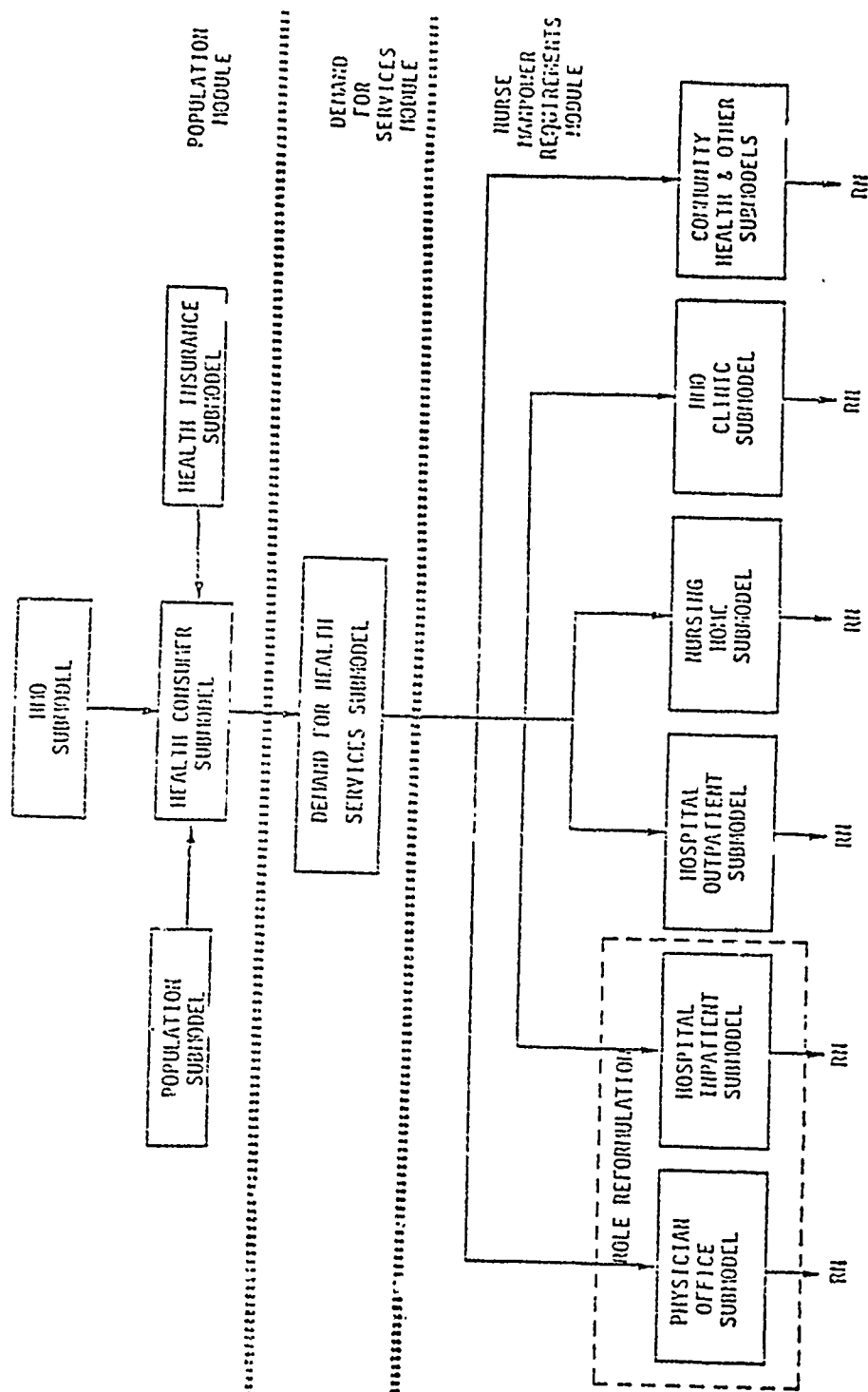


FIGURE 1. THE STRUCTURE OF THE VECTOR MODEL.

module to produce forecasts of the future U.S. population which characterize cohorts in terms of health insurance coverage and HMO enrollment. The first two types of data were obtained from the U.S. Bureau of census,^{1/} and were combined to produce population projections in terms of age/sex/family status/family income cohorts. The third type of data describes the population in terms of the fraction of each cohort eligible for benefits under the types of health insurance plans being considered. The fourth type of data describes the HMO formation rate, which is the number of HMOs that become operational in each year

The demand-for-services module requires as input projected per capita health service demands for each of the population cohorts projected by the population module. These input data consist of sets of per capita demands which characterize the consumption of health services by individuals enrolled in HMOs, and by individuals covered by different types of health insurance plans. Estimates of the per capita demands of HMO enrollees for the hospital inpatient and HMO clinic settings are assumed to be proportional to the per capita demands of their non-HMO enrollee cohorts. For other settings, the demands of HMO enrollees are assumed to be the same as for the general population.

The effect of a national health insurance plan on per capita demands is determined by estimating the change in the price of health services to the consumer resulting from implementation of the plan. The measure of price that is used is the effective coinsurance rate (i.e., the average fraction of expenses paid out-of-pocket by the consumer). Utilizing the results of previous studies on the price elasti-

^{1/} U.S. Bureau of Census, "Population Estimates & Projections," Current Population Reports, Series p-25, No. 541, February 1975; and U.S. Bureau of Census, "Income in 1973 of Families and Persons in the United States," Current Population Reports, Series p-60, No. 97, January 1975.

city of demand for health services,^{1/} the effect of changes in the coinsurance rate (i.e., the type of plan) on health service demands is then computed by the model.

By applying the appropriate set of per capita demands to the projections from the population module, the Vector model simulates the effects of changes in the population's health insurance coverage and in the number of HMO enrollees on health service demands in each provider setting. These health service demands are then used by the nurse manpower requirements module to estimate the impact of these two health system changes on requirements for nurses.

The principal objective of the Vector model is to evaluate the impacts of three anticipated health system changes on future requirements for nurses. These changes involved the implementation of National Health Insurance, expanded enrollments in Health Maintenance Organizations, and the reformulation of nursing roles. Therefore, the Vector model was specifically designed to enable model users to simulate a variety of alternative scenarios involving one or more of these changes, and to estimate the resulting impact on nursing requirements. Projections of the nurse manpower requirements without the above policy impacts are therefore, the secondary purpose of the model. The description of the model above shows that a sophisticated model often uses the manpower-population ratio as a starting point.

Note that the model starts with per capita demand for health services and ends up with a series of modifications to it according to the demographic and economic characteristics of population, and the extent of health insurance coverage and HMO formation.

1/ Heaney, Charles T. and Riedel, Donald C., From Indemnity to Full Coverage: Changes in Hospital Utilization, The Blue Cross Association, Chicago, 1970; and Newhouse, J.P. and Phelps, C.E., Price and Income Elasticities for Medical Care Services, presented at a Conference of the International Economics Association, Tokyo, Japan, April 1973. For the hospital outpatient and physician office settings the data from Scitovsky, Anne A. and Snyder, N. M., "Effect of Coinsurance on Use of Physician Services," Social Security Bulletin, June 1972 are used

SERVICE TARGET APPROACH

Manpower requirements are determined by the need of and the demand for the services provided by manpower, that is, the demand for labor is derived from the demand for its final product. Recognizing this fact, the service targets approach focuses on the factors affecting the demand for services and manpower productivity. This method is a policy oriented normative approach to manpower planning. To use this method, one sets the target of the types and amount of services "required" by the population of the area chosen. What types of and how much services are required are determined by the planner according to his judgement and information on the incidence of conditions requiring health services.

Once the service targets are established, manpower requirements are derived by applying factors related to manpower staffing and productivity. Manpower staffing practices are analyzed by means of task analysis, and this entails a detailed study of job performance in which job functions are identified and separated into tasks. This forms a basis for estimating manpower requirements for each occupation.

The service target method can best be illustrated by an example. First, the service target should be established as one ambulance per 10,000 people, for instance. Such target should be validated by experts. Next, the staffing pattern should be determined, say, as two emergency medical technicians per ambulance. Third, the productivity of each EMP should be decided, say as 40-hour week, measuring the output by the hours the technicians are available to respond to emergency calls. Then, the manpower requirements may be calculated as 41 men, assuming that each ambulance will operate round-the-clock, a 168-hour week. (Dividing 168 hours by a 40-hour week for each man equals four men. Multiplying by two men per shift equals eight men, which, when multiplied by five ambulances, equals 40 men. Adding one additional EMT as a relief man gives the community's total manpower requirements for EMT's 41 men.)

The weakness of this approach is that its validity entirely depends on the judgement of experts as to what types and amount of services are required for a

community. The problem presented here is similar to that which existed in using manpower population ratio, where as you may recall, the validity of the method depended entirely on how the ratio was used. The Delphi method may be used to achieve a consensus of the experts in determining the service target as it may be used in calculating the required manpower-population ratio. Of the four manpower models examined in this paper, the WICHE ^{1/} model uses the service target method as a principal input.

WICHE model: As the Vector model was described in some details as an example to show how the manpower-population ratio was used in a manpower model, in the following, the WICHE model is described in some details as an example to show how the service target method may be used in a manpower model.

The WICHE model differs from the other health manpower models in that it does not define structural relationships between the selected model variables. Rather, it presents variables to be included in the model and it guides the construction and estimation of relationships among them. The guides which are presented consist of past data on relevant relationships and the identification of various factors affecting them.

The structural relationship is to be defined by the "assumptions" or estimates which are to be obtained from a panel of experts according to the process provided in the model on the basis of the background data and information provided. The assumptions are to be estimated according to a consensus approach where all panel members agree on the estimates in the course of mutual consultation and discussion. Since these panel members are likely to be those who will be involved in planning, the model builders consider this process of obtaining the structural relationship important. Note that all panel members are expected to bring diverse perspectives to the process.

1/ Analysis and Planning for Improved Distribution of Nursing Personnel and Services: State Model.
Western Interstate Commission of Higher Education (WICHE). Boulder, Colorado

The WICHE nursing requirements projections involve: (1) population projections by age group, race/ethnicity-income specific illness prevalence rates; (3) projections of the quantity and types of health services required to treat the projected prevalence of ailments; and (4) projections of the number of nurses required to provide the needed health services.

Figure 2 shows my interpretation of the conceptual structure of the WICHE model as it pertains to the requirements projections. First, the projections of state populations by age, race/ethnicity, and income are based on U.S. Bureau of Census projections. Second, given the projected composition of the population according to the above groupings, the prevalence rate of ailments of the state population is estimated from past data on the basis of the judgments of experts in epidemiology. Third, given the estimated prevalence rates of ailments of the population, the quantity and types of health services required for treatment is projected from data on the disease-specific utilization rates of health services. The conversion of the estimated prevalence of illnesses into required health services is also based on professional judgment. Finally, given the projected quantity and types of health services required, the required number of nursing personnel is projected for each job setting and level of educational preparation. These last projections are also made on the basis of historical staffing pattern and data on the educational attainment of nurses by using professional judgment.

There are a few important assumptions made in the WICHE model development. Criticism of these assumptions illustrates the shortcomings of the service target method in general. The first assumption is that the quantity and types of health services required depend on the prevalence rates of ailments of age-race/ethnicity-income specific populations groups. Variations in the quantity and types of health services sought among different socioeconomic groups given the same prevalence of ailments are not considered by the model. In this model, socioeconomic characteristics of the population influence the quantity and types of health services required through their impacts on the prevalence rates of ailments for the population, but do not influence needed health services through impacts on the behavior or

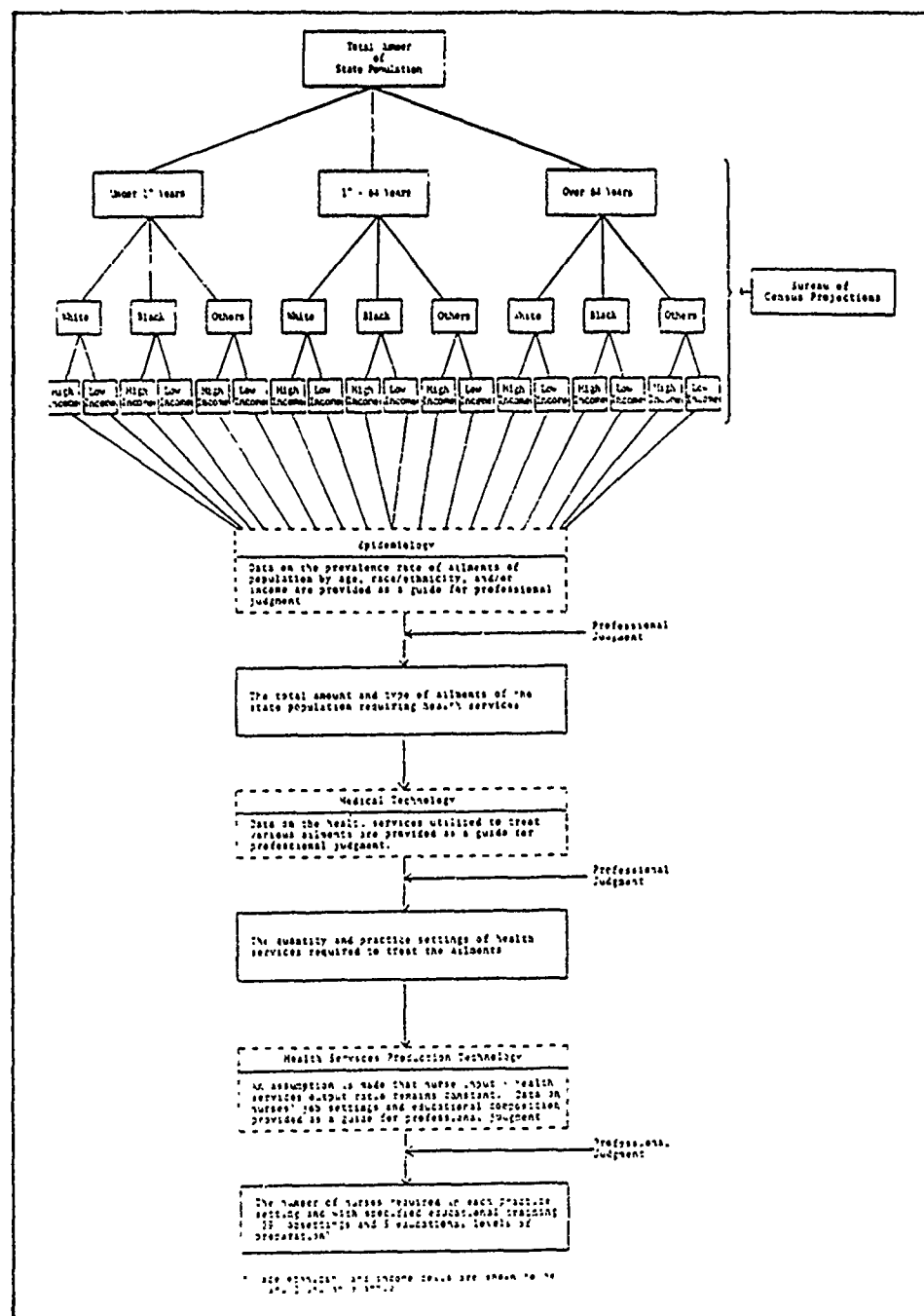


FIGURE 2. PROCESS OF MAKING NURSING REQUIREMENTS PROJECTIONS
WICHE MODEL AS INTERPRETED BY Kong-Kyun Ro

attitudes regarding seeking health services for a given illness. However, the expert panel can modify this by bringing the behavioral aspects of consumers with different socioeconomic backgrounds into the model.

Secondly, the model assumes that there is a constant or stable relationship between a specific disease or illness and the amount and types of health services required for treatment. The impacts of possible changes in medical science on the health services required to treat a given illness in the one-to-five years future are not explicitly incorporated into the model. The panel can, however, easily remedy this omission by considering the potential impacts of the advancements in medical science on the nurse labor input required.

NEED APPROACH

The need approach uses the health status of a given community as the starting point and estimates manpower requirements on the basis of the care needed to attain and maintain "good health." What constitute "good health" and the proper care to attain and maintain are obviously subject to definition. The definition of good health and standards of good medical care are to be set by professionals in the field.^{1/} Then, with quantitative information on the size of population and its epidemiological characteristics, one can define the health needs of a community, and on the basis of it, the manpower required to provide the proper services to meet the need. Seen in this way, the health needs approach may be viewed as an extension of the service targets approach in which the targets are set by the biological needs of the community, as professionally determined. Thus, the WICHE model may be described as a model using the need approach as well.

1/ Schonfeld, H.K., et al., "The development of standards for the audit and planning of medical care: Good pediatric care program content and method of estimating needed personnel," American Journal of Public Health, November 1968.

The distinctive character of the need method may better be illustrated by going back to the example of determining how many emergency medical technicians are required. First the health needs are prescribed by the opinions of experts as to the standard of medical care required by the critically injured, heart attack and stroke victims, and other emergency cases. These standard may stipulate that physician and technician services should be available 24-hour-a-day, to provide professional emergency care within 25 minutes of call, for an ambulance trip of no more than one hour within a geographic radius of 50 miles, employing a two-way radio-communication system and hospital based ambulances.

Then, the number of ambulances necessary to meet these standards of need can be calculated by taking into account the community's population characteristics, geography, travel conditions, location of hospital emergency rooms and emergency equipment, and the numbers and types of accidents that require emergency treatment within a certain time period. Given the above information, the number of EMT's required are derived in much the same way as in the service target approach.

The advantage of the need approach for a manpower planner is its logical basis in "what ought to be" as the reference point. Its weak point is the definitional problems, data and computational requirements. As pointed out before, expert opinions differ as to the definition of good health and standard care to attain and maintain it. In order to quantify the needs, services and manpower required, this approach requires detailed disaggregated data and sophisticated computational techniques.

ECONOMIC DEMAND APPROACH

It focuses on the "effective" demand for health care - willingness and ability of consumers to pay for health services - as the basic determinants of the demand for health services. In other words, the demand for manpower is derived from the demand for its final product and manpower requirements are determined by the demand for it. According to the traditional text book version, the demand for a good or service depends on relative price, (permanent) income, and tastes. Recently, the demand for health services has been analyzed in a new framework of analysis where health

services are treated as an input in the production of health, as manpower is an input in the production of health services. The new approach indicates that the shadow price of good health depends on the price of medical care, the value of time, and the efficiency of the production process. This means that in estimating the demand for health services, the value of time and educational level of consumers become the additional factors to consider.^{1/} Once the demand for health services in a community is estimated, manpower requirements may be determined on the basis of productivity estimates.

It should be noted here that the demand approach for projecting manpower requirements in the health field encounters an additional problem. It is the idea that the demand for health service is largely supplier determined in contrast to other goods and services where the demand for a product is determined by consumers.^{2/} It has an intriguing implications for making projections of manpower requirements in that the requirements increase as the supply increases because the demand for the final product, i.e., health services, increase as the number of suppliers increases. None of the approach examined in this paper, however, includes this supplier created demand effect in their methods of projecting the manpower requirements.

^{1/} Michael Grossman, "On the Concept of Health Capital and the Demand for Health," Journal of Political Economy, Vol. 80, No. 2, (March-April 1972), pp. 223-56. and Richard Auster and Kong-Kyun Ro, "The 'New Approach' and the Demand for Hospital Care," The Korean Economic Review, Vol. 25, November 1977.

^{2/} For discussions of the supplier-influenced demand for health services and its implications, see Uwe E. Reinhardt, Physician Productivity and the Demand for Health Manpower: An Economic Analysis, (Cambridge, Mass., Ballinger Publishing Co., 1975); Victor Fuchs and Marcia J. Kramer, Determinants of Expenditures for Physicians' Services in the United States, 1948-1968. Washington, D.C.: DHEW Pub. No. (HSM) 73-3013, December 1972; and Kong-Kyun Ro and John A. Powr Powers, "A Behavior Model of Physicians within the Framework of Leisure Versus Output Theory of Labor Supply," The Korean Economic Review, Vol. 24, December 1976.

The demand approach may be illustrated by using the example of the emergency medical technician once again. First, the demand for emergency ambulance services should be estimated. Such estimation may be made by studying the income level of the community, the price of an ambulance call, the size of population, etc. Alternatively, the demand estimate may be made by studying the utilization records of the past and then adjusting it according to the expected changes in the factors affecting the demand for ambulance services at the present. Another approach to estimating the demand is to use the estimate of the operators of ambulance services as a principal input. The conversion of the demand for ambulances into the demand for EMT's would proceed, as it has been described in the services targets approaches, based on data on staffing patterns and productivity.

There are many techniques to estimate the demand for health services and convert it to that of the demand for manpower. The most often used method is to use regression analysis where the independent variables consist of those factors hypothesized to influence the demand for health services according to the economic theory. One of the four models examined here, the CSF model 1/, uses this technique. Another model, the Pugh-Roberts model 2/, also uses a modified version of the demand approach as a part of its overall structure, which is System Dynamic approach.

The principal advantage of the demand approach is that it relies on the market to determine manpower requirements. Ideally, in a smoothly functioning market system, there will be no need for policy makers to estimate manpower requirements. The market will see to it that manpower requirements are met automatically through "invisible hand" of Adam Smith.

1/ A Micro Model for Assessing Nursing Manpower Demand and Supply. Community Systems Foundation, Ltd. (CSF) Washington, D.C.

2/ A National Model of the Supply of and Demand for Distribution of Nursing Personnel and Services. Pugh-Roberts Associates, Inc. Cambridge, Massachusetts.

In a community with an imperfect labor and health care market, however, manpower planning is essential. Therefore, there is a need for estimating manpower requirements. The demand approach is based on the recognition of this reality, and yet, attempts to base manpower planning on the dictations of the market. Even if one is prompted to perform manpower planning according to a normative dictate, it is still useful to estimate the market demand conditions and attempts to adjust the demand and supply conditions according to a given normative dictate, rather than ignore the market forces.

CSF Model : Another nurses' manpower model, namely CSF model, is described below as an example of how the demand approach may be used to project manpower requirements.

The CSF model consists of three types of submodels: (1) a set of submodels which forecast health services demand in acute care and long-term care settings; (2) two submodels for forecasting nursing requirements in ambulatory care, and community and public health care settings; and (3) a submodel for producing nurse supply projections. The health service utilization submodels were constructed using multiple regression procedures, the most frequently used technique in the demand approach. Each submodel is a single equation that predicts the demand for health services in each type of institution in terms of variables of two types: (1) demographic and economic characteristics of the population served by the institution; and (2) health system characteristics of the area in which the institution is located. Step-wise multiple regression procedures are used to select from a list of potential independent variables those that best "explain" the demand for health services.

The reliability of the projection made with a manpower model such as the CSF model depends on the explanatory power of the regression model used to forecast the demand for health services. This means that the independent variables should be selected with extreme care. In this respects the CSF model has not done well in the selection of the independent variables or/and the data available was of poor equality. The CSF model's "goodness of fit" of the individual regression equations, i.e. R^2 values for the twenty acute care regression equations and the eight long-term care regression equations indicate that, for

many of these equations, the independent variables account for only a small proportion of the variation in the dependent variables (i.e., patient days of care). It was, therefore, expected that these regression equations would provide only gross estimates of future nursing demand at the institutional level.

Another point to consider in using a manpower model based on the demand approach is how to convert the forecasts into estimates of future nursing demand. The CSF model uses the nurse staffing ratio as is done in the service target and need approaches. Two points should be made concerning these staffing ratios as used in the CSF model. First of all, these ratios are input data which must be provided by the model user. The user's manual accompanying the CSF model includes estimates of RN, LPN, and Aide staffing ratios which are classified according to the type and size of the institution and the region of the country so that the model user can select the appropriate staffing ratios for the particular institution being modeled.

A final consideration in using the CSF model involves the availability of historical data for the independent variables used in the regression equations. The model builders found that, on the basis of their experience during the testing and evaluation phase of the project, only limited amounts of historical data for the independent variables were available at the substate level. They, therefore, concluded that only simple linear projections of the independent variables were possible. While linear projections based on only a few historical data points may accurately predict trends in Census data, this may not be the case for data that describe future characteristics of the health care system of the substate area.

Pugh-Roberts Model : As mentioned before, the Pugh-Roberts model also uses the demand approach to make projections of manpower requirements. Instead of regression analysis, however, it uses the System Dynamics Approach. The System Dynamics models are variants of the operation research type models and consist of an initial set of variables whose values are called "level" variables and, while it is tempting to refer to them as stock variables, in reality they may represent flows. To each of these variables is then applied a rate of change which can be, in turn,

a function of other model variables. In essence, then, the System Dynamics model is a series of differential equations. The solution of these equations in terms of equation specification and parameter value assignment is handled by a software package called DYNAMO.

The Pugh-Roberts model consists of four sectors: nurse education, nurse employment, demand, and demographic sector. The nurse education sector forecasts the number of yearly graduates from each program. The nurse employment sector makes projections of the number of nurses employed in each employment setting. The demand sector forecasts the demand for nurses based on the projection of the demand for health services. Finally, the demographic sector represents key demographic characteristics of total and nurse population and how they impact on other sectors of the model.

The Pugh-Roberts model uses the demand approach in two sectors. First, the demographic sector projects the demand for health services based on the estimate of the illness incidence of the population. Second, based on the projection of the demand for health services, the demand sector estimates the demand for nurses using the "multipliers" for various factors chosen as influencing the demand for nurses, given the demand for health services.

Specifically, the demographic sector of the Pugh-Roberts model estimates the nation's population in terms of both total size and the distribution among ten age categories, as well as the fertility rates and death rates for members of each age category. The number and age distribution of those immigrating into the country is also included in the model.

Given the ten age groups, the rates of change of the population (per month) due to births, deaths, and immigration for each age category are derived from the "Series E" census projections. The rate of immigration is held constant at 400,000 people per year, and is distributed among various age categories. The model automatically "ages" the population each month by moving a fraction of each age group to the next-older age group. Female death rates, which are somewhat different from those reported for the total population, are used in computing the rates of deaths of nurses in each age category for each educational level.

The objectives of demographic sector are two fold. One is to provide a basis to calculate the age-specific incidence of illness in the population and, thereby, the magnitude of the population's need for health services. This enables the model to project the demand for nurses in the demand sector. The other is to provide a basis for projecting the number of applicants for each nursing education program. This, in turn, provides a basis for projecting the supply of nurses from the nursing education sector.

The Pugh-Roberts model determines the demand for nurses at each educational level, for each of the seven major employment settings in its "demand sector." Factors that affect the demand for nurses differ somewhat according to the employment settings. Since the majority of nurses are employed in hospitals, and because a set of factors that influence the demand for nurses in hospitals approximates the general set of factors that influence demand in most of employment settings, the determinants of the demand for nurses in hospitals are discussed here as representing the demand sector of the Pugh-Roberts model.

Factors affecting the demand for nurses in hospitals, as represented by the number of jobs available in hospitals, may be expressed as:

$$NEDH = f(WH, FSH, CBH, BNRH, INPNH, TCEFH)$$

where :

- NEDH = number of nurses at each educational level demanded in hospitals;
- WH = nurses' wages, represented by the ratio of average wages of nurses in hospitals adjusted for inflation to 1972 average ages;
- FSH = hospital's financial situation, as represented by an index measuring the degree of difficulty hospitals have in passing along increased costs;
- CBH = collective bargaining, as represented by the fraction of hospitals with collective bargaining agreements;

- BNRH - breadth of nurses' responsibilities in hospitals, as represented by a dimensionless index for which a value of 1.0 represents the breadth of responsibilities of nurses in 1972;
- INPNH - average intensity of patient needs for nursing services as represented by a combined index reflecting average length of stay and pre-admission screening; and
- TCEFH - technological change and other exogenous factors.

According to the values assigned to the multipliers, the Task Group considered the intensity of patient needs for nursing services, and nursing wages, as the two most important factors influencing the demand for nurses in hospitals. It projected a 20 percent decrease in the number of jobs available in hospitals with a 40 percent increase in nursing wages. On the other hand, the Task Group projected a decrease in the demand for nurses of five percent with an increase in the patient need intensity index of ten percent. It is also interesting to note that the Task Group expected technological changes and other exogenous factors to increase the demand for nurses by 20 percent by 1992.

It should be noted here that the concept of manpower requirements used by the Pugh-Roberts model corresponds to excess demand, that is, the difference between the manpower demanded and supplied at the prevailing wage rate. Figure 3 illustrates graphically the concept of manpower requirements as used by the Pugh-Roberts model and how the model projects the number of nurses demanded and that supplied.

As shown above, the CSF model uses the multiple regressions based on an implicitly formulated demand model and lets the actual data estimate the demand parameters, although the goodness of fit was not particularly good. The Pugh-Roberts model, on the other hand, comes up with the multipliers which correspond to the demand parameters based on the expert opinions. Which method is better has been the central point of the dispute between economists and System Dynamics people headed by Forrester and need not be elaborated here again.

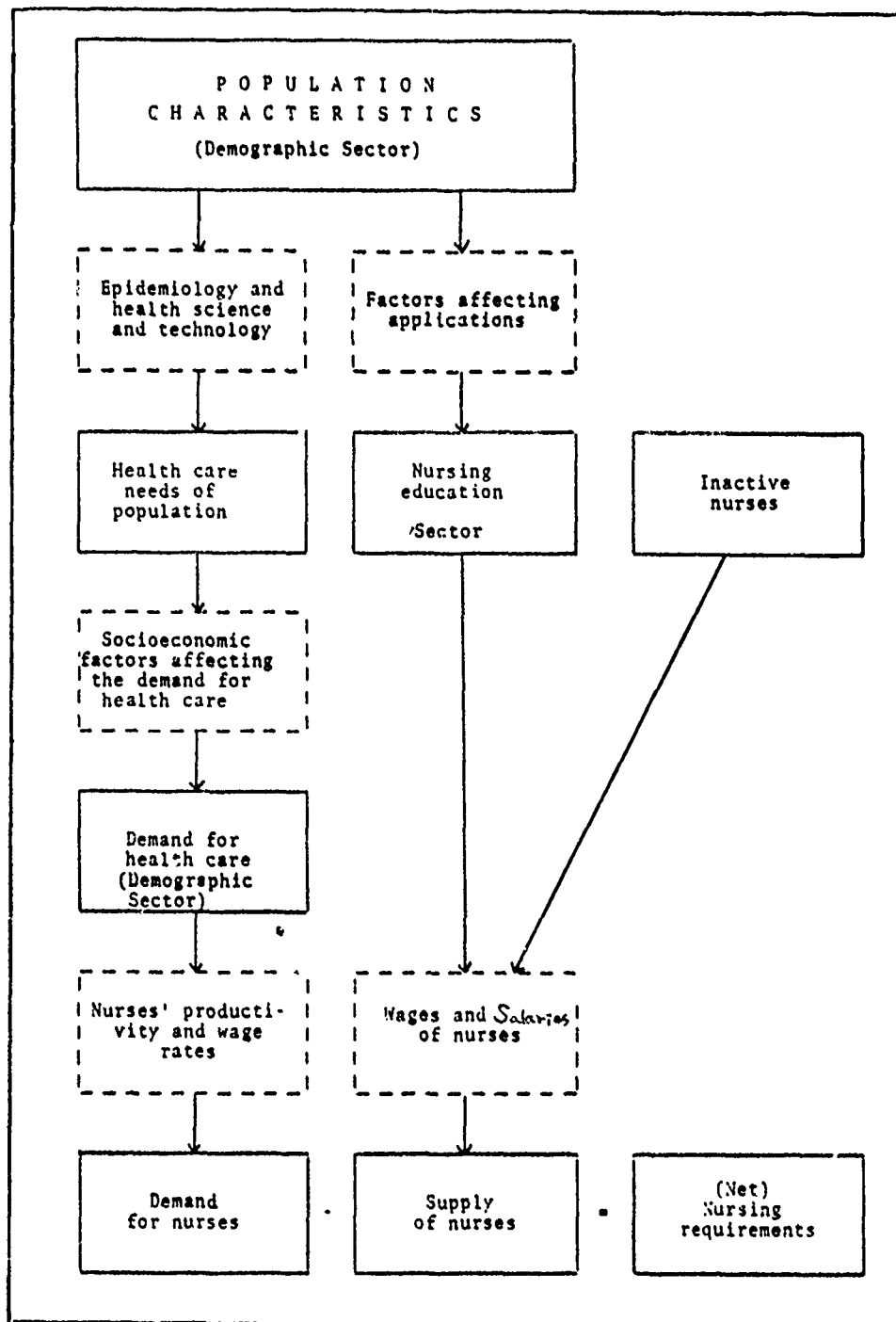


FIGURE 3 : A SCHEMATIC VIEW OF THE REQUIREMENT CONCEPT OF THE PUGH-ROBERTS MODEL AS INTERPRETED AND SIMPLIFIED BY Keng-Kyun Ro

Summary of the Four Approaches

Four distinct methods of estimating manpower requirements discussed above may be summarized as follows:

1. Manpower-population ratios (Ratio method) : This method involves the identification of a suitable manpower-population ratio for a future point in time and then the application of this ratio to the projected population to derive manpower requirements.
2. Service Target Approach (normative approach) : This method emphasizes the development of detailed standards for the provision of different kinds of services. Then, the standards are used to derive targets for the production of these services. Staffing and productivity standards are then converted into manpower required to attain such standards.
3. Health needs (or biologic needs) : This method seeks to determine what kinds, amounts and quality levels of services are required to attain and maintain a healthy population. The basis of the decision will be expert opinion and the data on health status and available technology. Then, service targets are converted into manpower requirements by means of staffing and productivity standards.
4. Demand approach : This method seeks to estimate the demand for health services based on the various factors theorized to influence the demand. One may use the traditional demand function where (permanent) income, relative price and taste are the determinants of the demand. Alternatively, one may use the new approach where health care is treated as an input for the production of health, and consumers demand health rather than health care.

The process underlying all the above manpower projection methods may be described briefly as follows: First, the factors affecting manpower supply and requirements have to be sought and selected. Second, how these factors affect the manpower supply and requirements, i.e., the structural relationship between the manpower requirements on one hand and the variables selected on the other, has to be specified. This indicates that the projections of the future manpower requirements also involve the projections of the magnitudes of the possible changes in those factors theorized to

affect the manpower requirements. In most cases, the structural relationship is assumed to remain constant. Finally, the manpower requirements must be projected to the target year on the basis of the expected changes in the variables chosen and the structural relationship estimated. 1/

Strengths and weaknesses of the above four methods may be summarized here. First, the manpower-population method is simple and easy to understand. Its strength, however, is also its weakness. It is too simplistic and may use inappropriate ratios that fail to take account of other important factors which determine the manpower requirements. As a starting point for estimating manpower requirements, this method provides a simple and readily available basis. Second method, the service target approach, recognizes that there are many other factors which determine the manpower requirements besides population size. The weakness of this approach is that it sets the service targets on a normative basis on more or less arbitrarily determined standards. Furthermore, in order to set the appropriate standards, it requires extensive data which are difficult to obtain.

The third approach, need method, is similar to the service target method and subject to the same kinds of criticisms. One has to decide what is needed in order to attain and maintain good health based on expert opinions. The definition of good health and what is needed to achieve it differ according to the expert. In fact, this method is identical to the service target method if the latter bases the target on the need concept.

1/ Harold M. Goldstein, "Methods of Projecting Supply and Demand in High Level Occupations," paper delivered at Annual Convention of the American Statistical Association, Philadelphia, Pennsylvania, U.S.A., September 8, 1965. Washington, D.C.: American Statistical Association, 1965.

Finally, the demand concept appears to me to be the most appropriate method to estimate the manpower requirements. This is because it relies on the market system rather than normative or expert opinions. This is not to say that this method does not have any weakness. The demand estimate is as good as the model and the data used. Its strength is that such estimate can improve as the model and the data used improves.

METHODOLOGIES TO ESTIMATE AND PROJECT MANPOWER SUPPLY:

Methods to estimate and project manpower supply differs according to the definition of supply. In the traditional labor economics, the supply of labor is defined as those who are currently employed and those who are unemployed but seeking employment. Those who are not actively seeking jobs but willing to work if the job opportunity improves and/or the wage level increases are in the labor force but do not constitute a part of the current manpower supply.

More appropriate definition would be to label those currently employed and actively seeking employment as the "active" supply and those are in the labor force and possess the requisite credentials but are not actively seeking jobs as the "inactive" or "potential" supply. In estimating the current supply of manpower, the accuracy and reliability of the estimate depends on the accuracy of the currently available data or the data collected. In projecting future supply of the manpower in the health field, most methods use one kind or another of the injection and leakage approach.

The leakages-injections approach in macro-economics looks at injections into and leakages from the circular flow of expenditures and incomes as the determinants of the level of Net National Products, that is, the GNP minus capital consumption. The greater the injections, which in macro-economics consist of domestic and foreign investment, and the smaller the leakages, which consist of savings and investment into foreign countries, the greater the NNP.

As applied to manpower projections, this approach postulates that the future supply demands on the estimated magnitude of injections of new graduates

of educational programs in health professions into the aggregate pool of health manpower and that of leakages of active health professionals from the pool. Leakages consist of (a) the emigration of active health professionals to foreign countries, (b) retirement and death, (c) involuntary unemployment or voluntarily being out of the labor force due to marriage or the lack of suitable employment opportunities at suitable locations. 1/

The WICHE model bases its projection of nurse manpower supply on the injection-leakage approach. Given the number of active nurses in the base year, the model makes a projection of supply by adding the estimated number of nurses expected to enter the supply pool, and subtracting the number of those expected to leave the pool. The inflows or injections into nurse supply are expected from three sources: (1) new graduates from nursing schools; (2) inactive nurses becoming active; and (3) immigrating nurses. The outflows or leakages are expected from four sources: (1) retirement, including those becoming temporarily inactive, and death; (2) suspension of licenses; (3) nurses returning to schools; and (4) out-migration. Figure 4 presents a graphical interpretation of the WICHE's injection-leakage model.

The supply projection by the WICHE model consists of presenting the relevant data and defining the factors affecting the rates of the inflows and the rates of the outflows of the types mentioned above. Estimation of the coefficients necessary to convert the current enrollment of nursing schools into the number of yearly graduates, the number of inactive nurses into the number of those returning to work, the number of active nurses into the number of those moving out of state, etc., are arrived at through the deliberations of the members of the expert panel.

1/ For a detailed explanation of the injection-leakage approach, see Neal Rosenthal, "Projections of Manpower Supply in a Specific Occupation," Monthly Labor Review, November 1966; and U.S. Department of Health, Education and Welfare, The Supply of Health Manpower, 1970 Profiles and Projections to 1990, DHEW Pub. No. (HRA) 75-38, Washington, D.C.: U.S. Government Printing Office, December 1974.

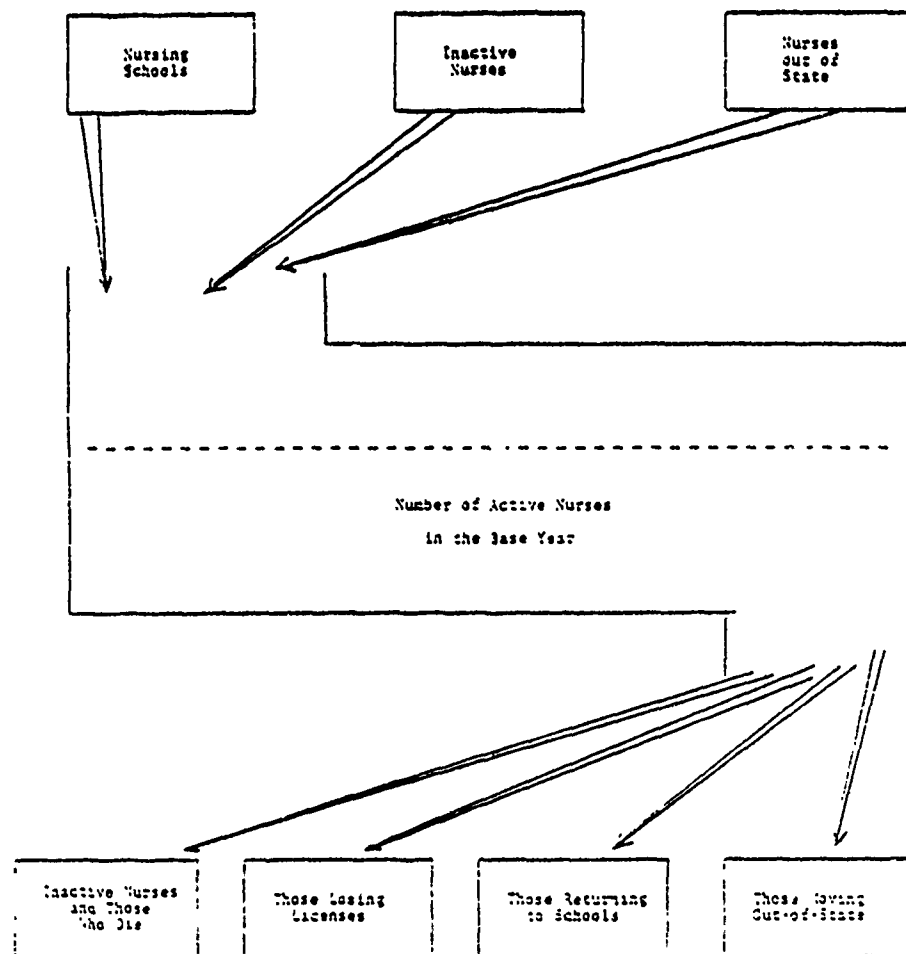


FIGURE 4 : INJECTION-LEAKAGE DIAGRAM FOR NURSING SUPPLY PROJECTIONS - WICHE MODEL AS INTERPRETED BY KONG-KYUN RO

The WICHE model makes an assumption that the technology of delivering health services exhibits a constant, or at least stable, relationship between the number of nurses employed and the quantity of health services produced. The possible impacts of future changes in labor-mix, nurses' productivity, nurses' wage rates vis-a-vis those of other inputs, and capital-labor ratios in the production of health services are not explicitly incorporated into the model. The model does not provide a guide as to how one may make professional judgments about the impacts of such changes on nursing requirements and supply.

The Pugh-Roberts model also uses a modified version of the injection-leakage approach in making projections of nursing manpower supply. Whereas the WICHE model simply defines the factors affecting the rate of the inflows and outflows and presents the relevant data, the Pugh-Roberts model not only selects these variables but also presents the multipliers which specify how these variables determine the rate of inflows and outflows. Figure 5 shows how the Pugh-Roberts model projects the supply of nurses. In the Pugh-Roberts model, the inflows and outflows consist of: (1) the number of licensed nurses including new graduates, who are employed or seeking employment, and who are willing to take available jobs at each job settings, (2) the number of nurses not employed or inactive but willing to take available jobs, (3) the number of nurses quitting to take other nursing positions, (4) the number of nurses retiring or becoming inactive, and (5) the number of nurses immigrating into the country. Determination of the above would enable model users to project the effective rate of nurses' supply at each point in time.

The Pugh-Roberts model's specification of the factors affecting the supply of nurses from both the nursing education sector and active nurses may be expressed in the following way:

$$FNWECE = f(PAJAS, W, BNR, EFOT)$$

where

$FNWECE$ = fraction of nurses at each educational level who are employed or are willing to consider employment;

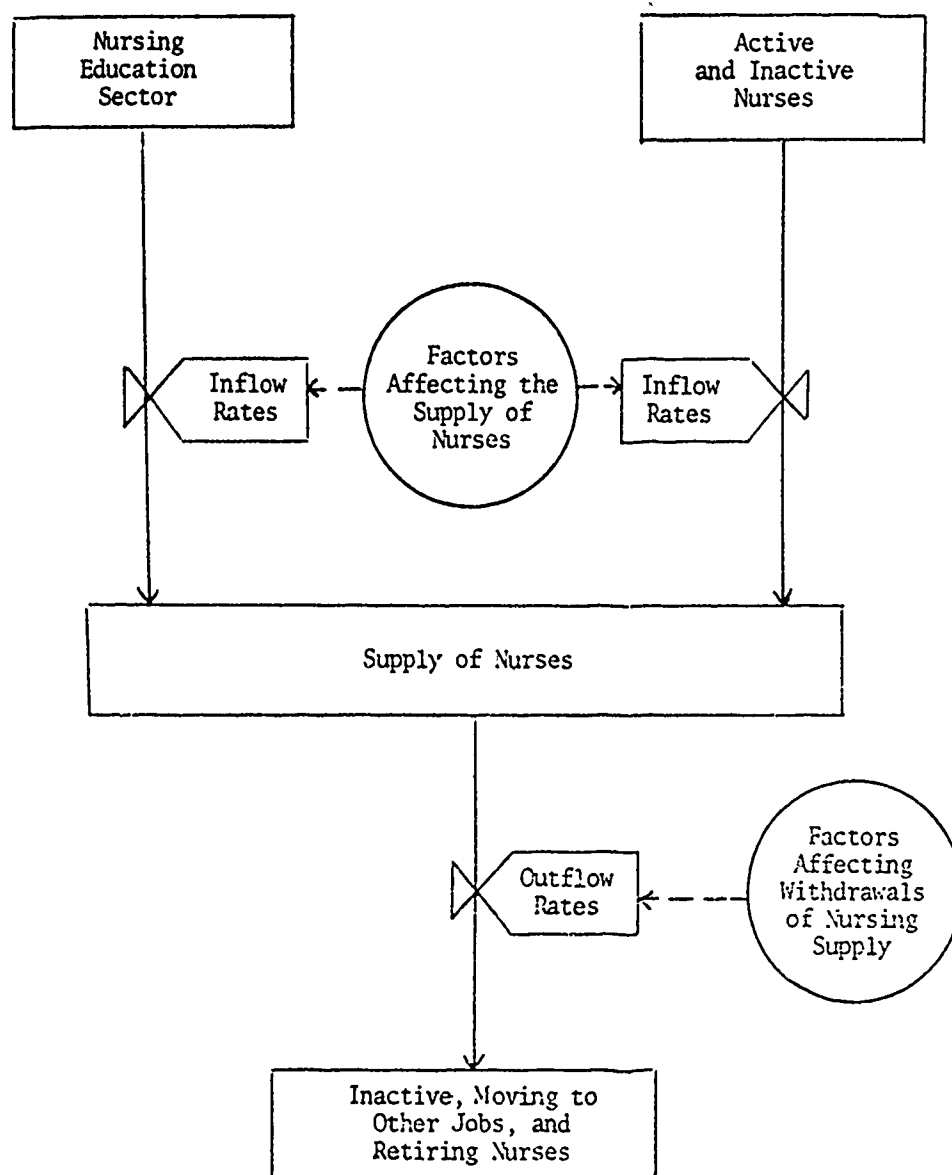


FIGURE 5: SCHEMATIC VIEW OF THE SUPPLY OF NURSES OF THE PUGH-ROBERTS MODEL AS INTERPRETED BY KONG-KYUN RO.

- PAJAS - perceived availability of jobs at each level, as presented by the ratio of the number of jobs available for nurses at each level to the total number of jobs for nurses at each level;
- W - nurses' wages, as represented by the ratio of wages of personnel at each level to 1972 wages for personnel at that level;
- BNR - breadth of nurses' responsibilities represented by a dimensionless index in which a value of 1.0 represents the breadth of nurses' responsibilities in 1972; and
- EFOT - changes in overall economy over time.

In terms of the values of impact multipliers given by the Task Group, by far the most important factor influencing the supply of nurses is nurses' wages. Other factors are judged to exert relatively insignificant impacts on the nurses' supply.

The CSF model uses multiple regressions to project the supply of nurses. The regression model, however, consists of defining the inflows and outflows of nurses into and out of the base year supply pool of licensed nurses. It does not specify the underlying causes which determine the rate of the addition to and the reduction from the existing manpower supply. The following equations show how the CSF model makes projection of nurse supply.

$$S(t) = \sum_{j=1}^n a_j(t) L_j(t) \quad (4.1)$$

$$L_j(t+1) = P_j L_j(t) + G_j(t) + (FN_j(t) - MO_j(t)) \quad (4.2)$$

where;

$S(t)$ - the number of nurses in year t (i.e., the active supply),

- $L_j(t)$ = the number of licensed nurses age j in year t ,
 $a_j(t)$ = the activity rate for nurses age j in year t (i.e., the ratio of employed nurses to licensed nurses),
 P_j = the one year probability of survival for nurses age j ,
 $G_j(t)$ = the number of new graduates age j licensed in year t ,
 $FN_j(t)$ = the number of new foreign, new endorsements, and reinstated nurses age j in year t , and
 $MO_j(t)$ = the number of nurses age j who migrate out of the area or whose licenses expire and are not renewed in year t .

It can be seen from equation (4:2) that the supply of licensed nurses is estimated on a yearly basis by taking account of new graduates, new foreign nurses, reinstated nurses, new endorsements, nurses who leave the area or whose licenses expire, and one-year survival probabilities. The active supply of nurses, equation (4:1) is then determined by multiplying the (age-specific) supply of licensed nurses by the (age-specific) activity rate. ^{1/}

An important factor which greatly influences the supply of health manpower is not explicitly included in any of the above models. It is the public policy towards the health of community. A policy decision might be made to enlarge or reduce the enrollment of educational institutions which train the manpower in the field. It might institute a health insurance system which would increase the demand for the health

^{1/} This model is based on an approach developed by Research Triangle Institute located at Chapel Hill, North Carolina, U.S.A. See Jones, D.C., et al., Procedure for Projecting Trends in Registered Nurse Supply, Research Triangle Institute, March 1975, (FR-24U-1024-2). Division of Nursing, Bureau of Health Planning and Resources Development, HRA, DHEW.

services which, in turn, would increase the demand for health professionals. 1/ Therefore a projection of the future manpower supply based on the injection-leakage method has to make assumptions about the future supply conditions which are determined by public policy or by other events.

As an example, let me cite the manpower supply projections my colleague and I made for Korea for 1980's. 2/ We made three different projections based on three different assumptions about the public policy which influences the future magnitude of injections and leakages. The assumptions behind projection I are that there will be no change in the public policy toward the size of enrollment of students in the schools of nursing and that of the emigration of nurses from Korea. Thus, it is assumed that (a) the amount of injections is the estimated number of new graduates based on the current enrollment of students in each class in educational programs in health professions, and (b) the amount of leakages remains at the current level in terms of number of health personnel leaving Korea and active professional participation.

Projection II assumes that there will be a change in the immigration policy toward foreign nurses in the U.S.A. but the Korean policy toward the size of nursing student body remains the same as Projection I. Thus, it assumes that the level of injections is the same as that assumed for Projection I but the level of leakages is smaller. Projection III assumes that the level of leakages is of the same magnitude as that estimated for Projection II, but the level of injections is higher. The assumption behind this projection is that the Ministry of Education increases student quotas of

1/ For a study of the impacts of alternative health insurance plans on manpower requirements, See Huang, Lien-fu and Elwood W. Shomo, "Assessment and Evaluation of Archetypal National Health Insurance Plans on U.S. Health Manpower Requirements," DHEW Pub. No. (HRA) 75-1, Rockville, Md., U.S. Department of Health, Education and Welfare, 1974.

2/ Kong-Kyun Ro and Mo-Im Kim, "Analysis of Health Resources in Korea," Report submitted to the U.S. Agency for International Development, 1975.

existing educational programs in health professions and/or increases the number of institutions authorized to offer such programs.

CONCLUDING REMARKS

Projections of manpower requirements and supply are made in order to avoid shortage or surpluses of the necessary manpower in the future. In a free and well functioning market system, there will be no need for such projections because price movements will eliminate any persisting shortages or surpluses. In most of the so-called mixed economy, however, there is a need for manpower planning and, therefore, manpower projections. There are two reasons for this. First, the concept of manpower requirements is subject to various definitions and interpretations. As mentioned in this paper, two of the four most commonly used method to estimate and project manpower requirements uses a normative definition of manpower requirements. In order to meet the manpower requirements as defined by a normative standard, one cannot rely on the market system to eliminate a possible shortage or surplus. Second, public policy in the mixed economy greatly influences the manpower requirements and supply, regardless of whether the policy is implemented through market system or not. This means that projections of manpower requirements and supply must be made which take account of the impacts of public policies. The Vector model is specifically built to evaluate the impacts of the national health insurance and Health Maintenance Organization formation. The Pugh-Roberts model also includes the impacts of public policies which influence the factors affecting the demand for and supply of nursing manpower.

One is bound to be bewildered by the variety of the methods used to make estimations and projections of manpower requirements. This is because the objectives and the concept of manpower requirements and supply are different among methods. For example, the Pugh-Roberts model uses the concept of manpower requirements which corresponds to the excess demand, whereas for other models, manpower "required" is what is demanded in the market or needed to meet a normatively set service target. The primary objective of evaluating such a variety of methods in this paper

has been to acquaint the readers with what is available and the special features of each method so that one may choose the best method on hand for his purpose. It should be pointed out that each method has relative strengths and weaknesses and different assumptions underlying the rationale for using a particular method. One should be particularly careful to be aware of assumptions made in the methods discussed in this paper. In presenting and evaluation the various methods, it is also hoped that researchers will improve the existing methods and invent new and better methods which will be needed in the future as the dictate of demand for such methods develop under changing circumstances in manpower planning.

MAN/COMPUTER INTERACTIVE TECHNIQUE
IN TRANSPORTATION SCHEDULING

PAUL L. TUAN

Transportation Operations and Information
Systems Department
SRI International
Menlo Park, CA 94025, U.S.A.

ABSTRACT. This paper gives some first-hand findings of research work conducted in vehicle and crew scheduling as applied to subway, bus, and airline operations. The experience of developing a man/computer interactive crew scheduling model and its implementation for one of the largest subway and bus companies in the U.S. is described. The process is essentially two phases -- (1) the automatic generation of a first-cut near-optimal solution; and (2) the refinement of the first-cut solution through an interactive procedure. Both phases involve techniques that require cooperation between scheduling logic and the scheduler's operational insights and human-factor considerations.

1. INTRODUCTION

This paper gives a capsule description of a man/computer interactive technique in transportation scheduling. Although, in general, a total transportation scheduling process involves three major segments - time table development, vehicle scheduling, and crew scheduling - for purpose of clarity this paper covers only the crew scheduling part. In the context of our research project, "scheduling" means the creation of crew work programs prior to schedule execution. It does not include the actual dispatch of crews and other dynamic elements that occur stochastically in the field. The process described herein is primarily designed and developed for urban transit operations, for example, subways and buses, although it was found that there is a close similarity between airline flight crew scheduling and ground transit operations scheduling. The same man/computer interactive technique is being utilized by both types of operations.

This paper is based on the work of a SRI research team which, besides the author, is comprised of Messrs. Jerome Johnson, David Marimont, Michael Tashker, and Bruce Robinson, all of whom are members of SRI's Transportation Operations and Information Systems Department.

2. THE CONCEPT OF MAN/COMPUTER INTERACTIVE PROCESS IN CREW SCHEDULING

Extensive research has been accomplished on various schemes of mathematical programming for seeking optimal or near-optimal solutions. The references listed at the end of this paper [1 - 11] are among the significant studies that have been conducted in the past. However, there has not been sufficient work to incorporate field operating experience, schedule makers' insight, and subtle work rules into the scheduling logic. Despite the desire of transit companies for more automation in all aspects of transportation operations, the importance of the perception, experience, and systems control of the schedule makers already has been stressed. A pure "black-box" approach in scheduling is normally unattractive to the scheduling personnel and unacceptable to the operating authorities.

As a corollary to the above, the scheduling system must be responsive to man by having on-line interactive capabilities. The interaction capability should have the following minimal requirements:

- o Short, on-line response time.
- o System prompting and tutorial facilities for the inexperienced user.
- o At least two levels of system control by the on-line user:
 - Macrolevel - The schedule maker can influence the outcome of the schedule through parametric control.
 - Microlevel -- The schedule maker can make arbitrary improvements subject to computer check and improvement quality indicators.
- o Immediate on-line feedbacks to let the user know of his performance.
- o On-line graphics, capability, including histograms, quick quality indicators, bar charts, system diagrams, and so forth.
- o CRT should have split-screen, roll, blinking, grey-tone variation capabilities.
- o High-speed on-line printing as well as off-line printing (as a backup).

Figure 1 provides an overview of the man-machine interactive scheduling process. The crew scheduling process begins with a new vehicle timetable; this new timetable is supplied as an input to the crew-scheduling process. The computer automatically generates a first-cut crew schedule from this vehicle schedule. On this first-cut crew schedule, the scheduler will build more efficient schedules by manipulating and perturbing various schedule parameters. As such, the automatic schedule generator need not generate perfect schedules, only reasonable schedules.

The first-cut schedule is modified by the scheduler instructing the computer to make certain desired modifications or refittings; this results in a modified schedule. The computer then evaluates the modified schedule by providing a detailed list of various measures of efficiency (or quality indicators). The modified schedule can be rejected outright if the modification has not resulted in improvement. If the schedule is acceptable in modified form, the computer produces reports and statistics and the finalized schedule. If not acceptable, the computer will provide schedule quality indicators that allow determination of where the schedule is inefficient or inappropriate. Using the feedback provided by

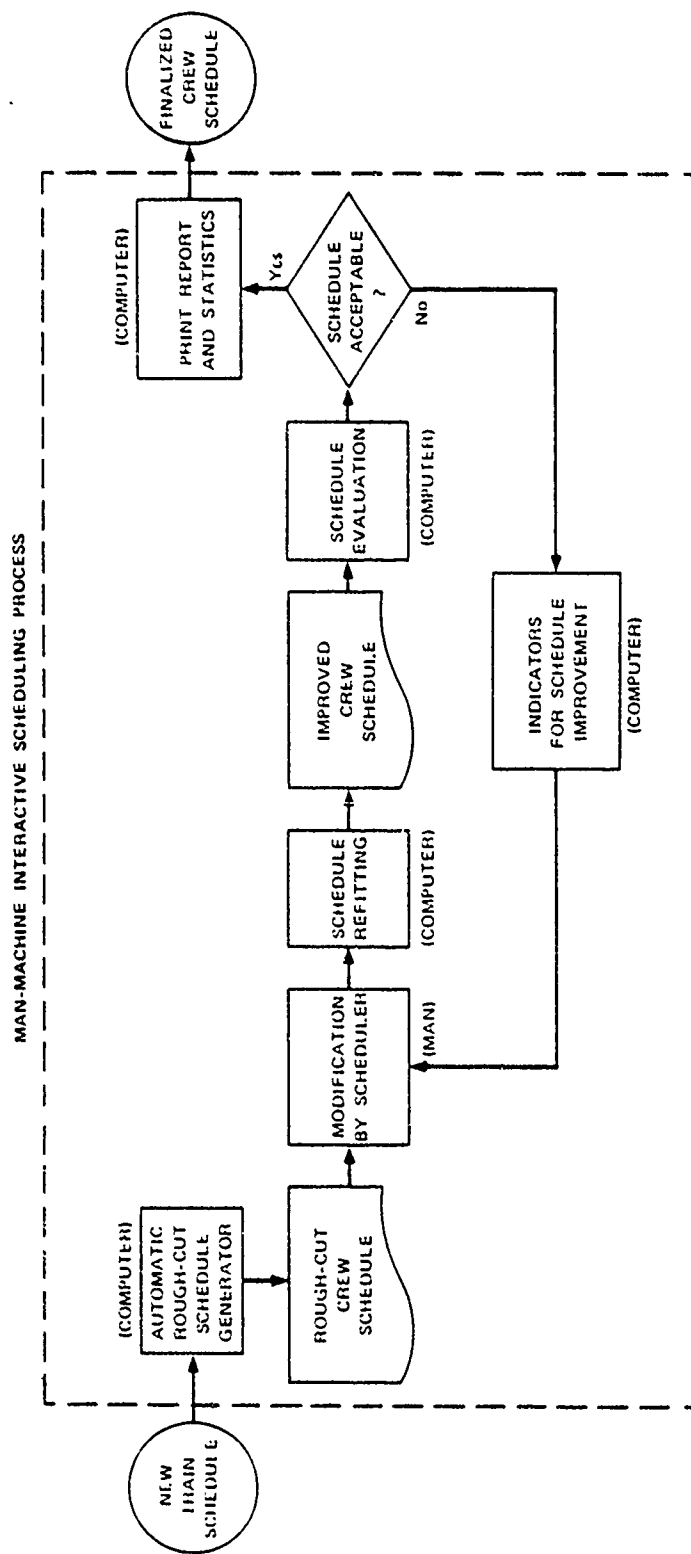


FIGURE 1 MAN-MACHINE INTERACTIVE SCHEDULING PROCESS

these indicators, the scheduler can make new modifications that result in another improved schedule, which in turn is evaluated for acceptability. This procedure is repeated until an acceptable schedule is found. Because the computer has taken over most of the routine operations, mechanical decisions, evaluation, and report and statistics generation, the scheduler can use his judgment and experience to quickly evaluate numerous schedules and to converge on an efficient schedule.

The interactive process for developing crew work programs can be viewed as having two levels -- the macrolevel and the microlevel. The macrolevel is also known as the "parametric level," which is represented by the first-cut phase. During this phase the schedule maker can vary the parameter values and control variables at the input stage to induce different sets of first-cut results. Please see Figure 2 and 3 for CRT images of input decision aids. With this approach, the schedule maker can successively improve his most desirable first-cut solution until the output is achieved.

During the second phase -- the refinement phase -- the schedule maker is given the facility to revise or construct crew runs at a very elementary level -- for example, move a revenue piece from run X to run Y, add a depot function at the beginning of run Z, or connect two one-trip runs into a two-trip run. This flexibility should enable the schedule maker to incorporate his experience, judgment, and insight into a work program at a microlevel.

3. FIRST-CUT PHASE

The key elements of the heuristic logic for the first-cut phase are contained in the following set of guidelines, listed in order of importance:

- (1) Each run is extended as far into a peak operating period as possible without violating workrule constraints. Conversely, no run should lie completely outside a peak period.
- (2) As many maximum-workpiece runs as possible are built, subject to workrule constraints. The number of these runs is estimated from train requirements for the line.
- (3) Relay breaks that fall during peak periods are minimized.

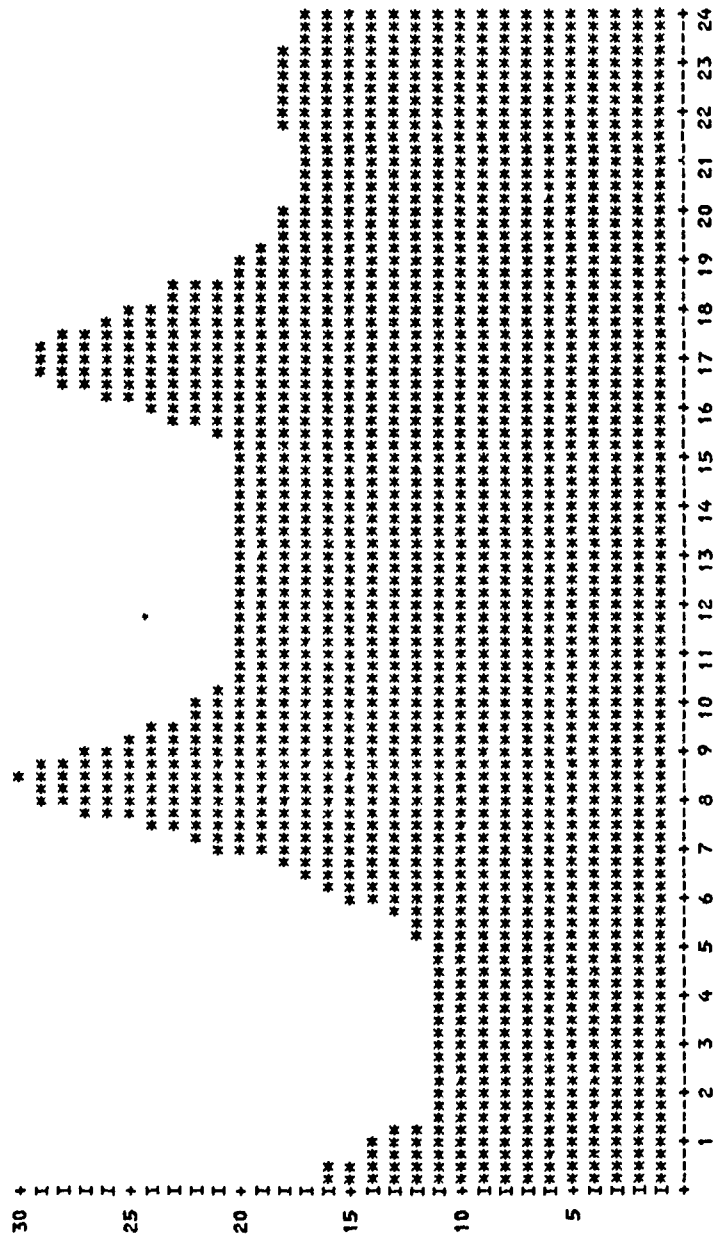


Figure 2 - TRAIN REQUIREMENTS (Y-AXIS) VERSUS TIME (X-AXIS)

```

1.  TITLE  *TEST RUN OF FRSTMD WITH 2A01 TIMETABLE *
2.  TITLE  *X2A01.RUF73,3EST,10M,6P01,9;NONPK BLW=20*
3.
4.
5.
6.
7.
8.
9.  SECTION *WORKRULE*
10. NREGT * 8: 0* MAXIMUM STRAIGHT TIME (HRS:MINS)
11. NOVLT * 0:59* MAXIMUM OVERTIME (HRS:MINS)
12. LUNCH * 0:35* MINIMUM LUNCH TIME (HRS:MINS)
13. LUNMIN * 3:20* MINIMUM TIME FROM REPORT TO LUNCH START (HRS:MINS)
14. LUNMAX * 5:40* MAXIMUM TIME FROM REPORT TO LUNCH END (HRS:MINS)
15. NCAB * 6: 0* MAXIMUM CAB TIME FOR SEQUENTIALS (HRS:MINS)
16. NREPT * 0:15* MINIMUM REPORT TIME (HRS:MINS)
17. MTCPUT * 1: 0* MINIMUM TC FOR PUT-INS AND LAY-UPS (HRS:MINS)
18. MTCADD * 0:30* MINIMUM TC FOR ADDS AND CUTS (HRS:MINS)
19. MTCMSC * 1:30* MINIMUM TC, MISCELLANEOUS (HRS:MINS)
20. NSPECT * 3: 0* MINIMUM ACTUAL TIME FOR SPECIALS (HRS:MINS)
21. END *****
22. SECTION * PAYROLL*
23. BPAYMOT * 8.330* BASE PAY, MOTORMEN ($/HR)
24. BPAYCON * 7.910* BASE PAY, CONDUCTORS ($/HR)
25. OVTRATE * 1.500* RATIO OF OVERTIME PAY TO STRAIGHT TIME PAY
26. RETRATE * 1.500* RATIO OF RETURN TIME PAY TO STRAIGHT TIME PAY
27. NDIFF1 * 2000* START OF NIGHT DIFFERENTIAL PAY PERIOD (MIL TIME)
28. NDIFF2 * 559* END OF NIGHT DIFFERENTIAL PAY PERIOD (MIL TIME)
29. MAXRADI * 5* MAXIMUM RADIO TIME (MINS)
30. END *****
31. SECTION * LINOBJ*
32. LINE * 2A* DATA BELOW FOR LINE 2A
33. AMPKST * 615* START OF AM PEAK PERIOD (MIL TIME)
34. AMPKSP * 1000* END OF AM PEAK PERIOD (MIL TIME)
35. PHPKST * 1545* START OF PM PEAK PERIOD (MIL TIME)
36. PHPKSP * 1830* END OF PM PEAK PERIOD (MIL TIME)
37. BEGTIM * 2400* BEGINNING OF FIRST CUT PROCESSING (MIL TIME)
38. FINTIM * 1200* END OF FIRST CUT PROCESSING (MIL TIME)
39. NBLOW * 4* NUMBER OF BLOWOT CARDS BELOW (MAX OF 40)
40. BLOWOT *1: 615-1000= 0/240*STATION:FROM(MIL TIME)-TO(MIL TIME)=MIN BLOWOT/M
41. BLOWOT *1:1001-1544= 20/240*STATION:FROM(MIL TIME)-TO(MIL TIME)=MIN BLOWOT/M
42. BLOWOT *1:1545-1830= 0/240*STATION:FROM(MIL TIME)-TO(MIL TIME)=MIN BLOWOT/M
43. BLOWOT *1:1831- 614= 20/240*STATION:FROM(MIL TIME)-TO(MIL TIME)=MIN BLOWOT/M
44. EXPAND * 15* BLOWOT EXPANSION % FOR STRETCHED RUNS
45. NDROP * 1* NUMBER OF DROPBACK CARDS BELOW (MAX OF 50)
46. DROPBK *9: 816- 817=W* STATION:FROM(MIL TIME)-TO(MIL TIME)=W,1,2 OR 3 TRA
47. EQSTGRP * 2* NUMBER OF EQUIVALENT STATION GROUPS (MAX OF 5)
48. EQST * 1- 2- 3- 0- 0* EQUIVALENT STATIONS (STATION NUMBERS)
49. EQST * 9- 0- 0- 0- 0* EQUIVALENT STATIONS (STATION NUMBERS)
50. NBRKPT * 2* NUMBER OF BREAKPOINTS (MAX OF 9)
51. BRKPT *2330- 1- 10- 6* BREAKPT(MIL TIME-STA. NO.-NO.RUNS MIDN-NO.RUNS PM
52. BRKPT * 30- 9- 10- 6* BREAKPT(MIL TIME-STA. NO.-NO.RUNS MIDN-NO.RUNS PM
53. PRUNMD * 4* PIECES PER MIDNIGHT STRETCH RUN
54. PRUNPM * 4* PIECES PER PM STRETCH RUN
55. END *****
56. SECTION * STADEF*
57. STATION *180TH ST= 3*
58. TROUND * 0* TURN-AROUND-OWN-TRAIN TIME (MINS)
59. REPORT *YES- 0* REPORT OK? (YES/NO-WHERE IF NO)
60. RELIEF *YES- 0* RELIEF OK? (YES/NO-WHERE IF NO)

```

FIGURE 3

A Partial List of Input Parameters to SRI's
Run-Cutting Program

- (4) Every partially developed run that is built into or through a peak period is completed as early as possible if there is no chance of it reaching later peak periods.

These guidelines are combined with a group of parameters calculated from the vehicle requirements function of the line to define a set of run types and a related set of parameters that can be used to form individual pieces of work from the timetable into a work program containing attributes desired by the schedule maker. Six such run types are currently defined:

- (1) Midnight Stretch Runs
- (2) P.M. Stretch Runs
- (3) A.M. Stretch Runs
- (4) A.M. Tandem Runs
- (5) P.M. Tandem Runs
- (6) One Trippers

Several parameters are currently being developed so that the schedule maker can override the "background" computer logic and specify a set of attributes for the final work program before a complete, first-cut computer run is performed. Some of these override features include:

- o Setting the size and identity of each of the first five run types described above.
- o Setting the timing and duration of relay breaks to accommodate put-ins, lay-ups, and the like.
- o Determining the overall directional imbalance within a run type.
- o Specifying first and last pieces for any run.

4. REFINEMENT PHASE

During this phase the schedule maker can review finished work programs (whether they are first-cut versions or refinement versions) via an on-line CRT and determine whether further improvements can be made. To serve as decision aids, the

computer also provides various quality indicators. See Figure 4 for a sample output of a first-cut solution. If it is decided that further improvements should be made, the schedule maker may instruct the computer, via an on-line terminal, to generate a finished work program with at least the same information content and format as the current manually generated report.

In the computer-aided system, the schedule maker will always receive a computer feedback (via on-line CRT or typewriter terminal) on the quality of the latest edition of the work program in progress. These indicators reflect the quality of either an automatically generated work program or a work program resulting from man-computer interaction. The schedule maker will determine whether there should therefore be additional improvement to continue the process or whether the process should be terminated since there is no more improvement to be made. There are four types of measures listed below in order of usefulness during schedule development:

- o Overall indicators
- o Trip profile distribution
- o Schedule coverage check
- o Time distribution.

The first three groups can be used as quick-and-dirty decision rules to measure whether a work program has reached a near-optimal resolution. The last group is more or less for analysis purposes once the work program is developed. Please see Figures 5 and 6 for samples of quality indicators.

5. COMPUTER UTILIZATION AND APPLICATION EXPERIENCE

The above system was developed on the Stanford University's IBM 370 system with remote-terminal time-sharing capabilities. Through a nation-wide telecommunication network the system has also been used from other locations in the United States via voice-grade lines, for example, New York City. System adaptation was also made by another time-sharing facility which uses remote CRT's with 4800 baud transfer rate and adjacent high speed printers. The response time per run (first-cut) at the terminal ranges from a few seconds to a few minutes dependent upon the total workload of the computer system. In almost all the cases the first-cut solution achieves an overall efficiency of over 90% as compared to the best solution that can be developed by an experienced schedule maker

2.
3.
4.
5.
6.
7.
8.
9.
10.
11.
12.
13.
14.
15.
16.
17.
18.
19.
20.
21.
22.
23.
24.
25.
26.
27.
28.
29.
30.
31.
32.
33.
34.
35.
36.
37.
38.
39.
40.
41.
42.
43.
44.
45.
46.
47.
48.
49.
50.
51.
52.
53.
54.
55.
56.
57.
58.
59.

Figure 4

PAGE: 1

TEST RUN OF FRSTMD WITH 2A01 TIMETABLE
X2A01.RUF73,BEST,10M,6P81,9;NONPK BLW=20
241ST ST= 1 238TH ST= 2 180TH ST= 3
NEW LOTS= 9

RUN	PT	R	REPT	L	LEAV	A	ARRV	L	LEAV	A	ARRV	R	RELF	L	LUND	LUNE	ACTL	ALLD	DIFF	RAD	RUN
101	4M	1	2329	/1	2344	9	111	/9	216	1	343										
				B1	519	9	644A	/9	702A	1	828	/1	828	1	343	519	8:59	9:29	6:30	5	101
102	4M	1	2344	/1	2359	9	126L	A9	236	1	403										
				B1	531P	9	656A	/9	712A	1	838	/1	838	1	403	531	8:54	9:21	6:15	5	102
103	4M	1		0	/1	15	9	142	/9	256	1	423									
				B1	542	9	708A	/9	728	1	855	/1	855	1	423	542	8:55	9:23	5:59	5	103
104	4M	1	20	/1	35	9	202	/9	316	1	443										
				D1	601A	9	728	/9	747	1	916	/1	916	1	443	518	8:53	9:24	5:39	5	104
105	4M	1	40	/1	55	9	222	/9	336	1	503										
				D1	617A	9	744	/9	805	1	935	/1	935	1	503	538	8:55	9:23	5:19	5	105
106	4M	1	100	/1	115	9	242	/9	356	1	523										
				B3	647P	9	759	/9	817	1	947	/1	947	3	537	647	8:47	9:11	4:59	5	106
107	4M	1	120	/1	135	9	302	/9	416	1	543A										
				B1	637A	9	806	/9	823	1	953	/1	953	1	543	637	8:33	8:50	4:39	5	107
108	4M	1	140	/1	155	9	322	/9	436	1	603A										
				B3	658P	9	812	/9	830	1	1000L	/1	1000	3	617	658	8:20	8:30	4:19	5	108
109	4F	1	400	/1	415	9	542A	A9	652A	1	817										
				B1	859	9	1025	/9	1052	1	1217	/1	1217	1	817	859	8:17	8:26	1:59	5	109
110	4M	9	26	/9	41	1	208	/1	235	9	402										
				D9	554P	1	719	/1	747	9	920	/9	920	9	402	437	8:54	9:21	5:33	5	110
111	4M	9	41	/9	56	1	223	/1	255	9	422										
				D9	616A	1	741	/3	816P	9	932L	/9	932	9	422	457	8:51	9:17	5:18	5	111
112	4M	9	101	/9	116	1	243	/1	315	9	442										
				D9	606A	1	731	/1	809	9	938	/9	938	9	442	517	8:37	8:56	4:58	5	112
113	4M	9	121	/9	136	1	303	/1	335	9	502										
				D9	629A	1	754	/1	819	9	946	/9	946	9	502	537	8:25	8:38	4:38	5	113
114	4M	9	141	/9	156	1	323	/1	355	9	522A										
				C9	641P	1	806	/1	829	9	955	/9	955	9	606	641	8:14	8:21	4:18	5	114
201	2F	1	420	/1	435	9	602A	A9	720A	1	846	B1	921	1	846	921	5:1	8:1	0	1:39	0 201
202	4F	1	436	/1	451	9	618A	A9	735P	1	903										
				B1	939	9	1105	/9	1132	1	1257	/1	1257	1	903	939	8:21	8:32	1:23	5	202

Output Format for SRT's Run-Cutting Program

270.
271.
272.
273.
274.
275.
276.
277.
278.
279.
280.
281.
282.
283.
284.
285.
286.
287.
288.

PAGE: 6

TEST RUN OF FRSTMD WITH 2A01 TIMETABLE
X2A01.RUF73.BEST.10M.6P01.9;NONPK BLW=20

PIECES PER RUN	OVERALL QUALITY INDICATORS (TIMES IN HRS:MINUTES)		
	NUMBER OF RUNS	ACTUAL TIME	OVER- TIME
1S	1	1:32	0: 0
2	14	69: 5	0: 0
3	1	7: 7	0:54
4	60	502:25	28:57
TOTAL	76	580: 9	29:51
			ALLD TIME
			4:38
			112: 0
			9:21
			523:41
			649:40

Figure 5

Quick Quality Indicators

57.
58.
59.
60.
61.
62.
63.
64.
65.
66.
67.
68.
69.
70.
71.
72.
73.
74.
75.
76.
77.
78.
79.
80.
81.
82.
83.
84.
85.
86.
87.
88.
89.
90.
91.
92.
93.
94.
95.
96.
97.
98.
99.
100.
101.
102.
103.
104.
105.
106.
107.
108.
109.
110.
111.
112.
113.
114.

PAGE: 2

TEST RUN OF FRSTMO WITH 2A01 TIME TABLE
X2A01.RUFF3.BEST.10H.6P81.9:NONPK BLN=20
RUFF3-GENERATED VERSION OF 1500. UNION

RUN	OLD	PT	REPORT	LUNCH	T/C	CAB	RELAY	ACTUAL	BOOST	ONTIME	O/TPEN	D/HPEN	ALLO	DIFFL	RADIO
307	307	48	0115	0139	01.0	5:51	0:49	7:34	0:26	01.0	01.0	01.0	81.0	2:34	01.0
308	308	28	0115	0135	1:53	2:55	01.0	5:30	2:22	01.0	01.0	01.0	81.0	31.6	01.0
309	309	48	0115	11.2	01.2	5:50	0:50	8:17	01.0	01.0	01.4	01.0	81.1	31.20	01.5
310	310	48	0115	11.4	01.0	5:55	0:50	8:12	01.0	01.0	01.6	01.0	81.1	31.32	01.5
311	311	48	0115	0135	0134	5:52	11.1	8:17	01.0	01.0	01.9	01.0	81.2	31.44	01.5
312	312	48	0115	0135	0139	5:53	11.0	8:22	01.0	01.0	01.11	01.0	81.3	31.56	01.5
313	313	28	0115	0135	1:40	2:57	01.0	5:27	2:33	01.0	01.0	01.0	81.0	31.41	01.0
314	314	48	0115	0135	1:6	5:53	0:43	8:32	01.0	01.0	01.76	01.0	81.4	4:20	01.5
315	315	4P	0115	0135	1:32	5:51	0:43	8:56	01.0	01.0	01.28	01.0	9:24	4:50	01.5
316	316	28	0115	0135	1:32	2:58	01.0	5:20	2:40	01.0	01.0	01.0	81.0	41.0	01.0
317	317	4P	0115	0135	1:34	5:50	0:42	8:59	01.0	01.0	01.28	01.0	9:24	5:12	01.5
318	318	4P	0115	0135	1:36	5:51	0:42	8:59	01.0	01.0	01.4	01.0	9:24	5:14	01.5
319	319	48	0115	0135	0:46	5:47	0:46	8:17	01.0	01.0	01.28	01.0	9:23	5:26	01.5
320	320	4P	0115	0135	1:30	5:49	0:46	8:55	01.0	01.0	01.30	01.0	9:29	5:36	01.5
321	321	4P	0115	0135	1:32	5:49	0:48	8:59	01.0	01.0	01.0	01.0	81.0	4:47	01.0
322	322	28	0115	0135	1:33	2:57	01.0	5:20	2:40	01.0	01.0	01.0	81.0	5:13	01.5
323	323	4P	0115	0135	1:36	5:48	0:46	8:58	01.0	01.0	01.29	01.0	81.0	5:13	01.5
324	324	28	0115	0135	1:22	2:54	01.0	5:16	2:54	01.0	01.0	01.0	81.0	81.0	01.0
325	325	28	0115	0135	01.0	2:54	01.0	4:23	3:37	01.0	01.0	01.0	81.0	81.0	01.0
326	326	48	0115	11.4	01.0	5:49	11.9	8:17	01.0	01.0	01.9	01.0	81.0	81.0	01.0
327	327	48	0115	11.0	01.0	5:49	0:32	8:19	01.0	01.0	01.10	01.0	81.0	81.0	01.0
328	328	48	0115	0136	0:52	5:48	0:52	8:23	01.0	01.0	01.12	01.0	81.0	81.0	01.0
329	329	48	0115	0136	0:55	5:49	0:40	8:15	01.0	01.0	01.6	01.0	81.0	81.0	01.0
330	330	48	0115	0142	01.0	5:47	0:46	7:30	01.0	01.0	01.0	01.0	81.0	81.0	01.0
331	331	4P	0115	0135	1:14	5:48	0:43	8:35	01.0	01.0	01.18	01.0	81.0	81.0	01.0
332	332	4P	0115	0135	1:18	5:49	0:44	8:41	01.0	01.0	01.21	01.0	91.2	4:20	01.5
333	333	4P	0115	1:53	01.0	5:49	0:50	8:47	01.0	01.0	01.24	01.0	91.1	4:32	01.5
334	334	4P	0115	0135	1:18	5:50	0:55	8:53	01.0	01.0	01.27	01.0	91.0	4:44	01.5
335	335	4P	0115	0135	1:20	5:48	11.0	8:58	01.0	01.0	01.29	01.0	91.0	4:56	01.5
336	336	4P	0115	11.1	1:35	5:46	0:15	8:52	01.0	01.0	01.26	01.0	91.0	7:42	01.5
TOTAL 191.0 57:59 49:16 393:18 60:36 5801.9 47:15 5:32 13:53 11:18 649:40 2081.5 41.0															

PAYROLL DATA

TYPE	RATE
MOT	\$ 8.33
CON	7.91
TOTAL	308.56 941.65 800.09 6387.19 984.14 9421.63 767.34 89.86 225.47 21.110550.58 3379.27 64.96

* TOTAL COST = \$ 13994.80 *

Figure 6
Time Distribution Analysis

Occasionally, the first-cut solution is equal or better than the best manual solution. With minor refinement the best solution can often be reached within minutes or hours as opposed to days or weeks under a pure manual system.

Research is currently underway to develop an interactive system that can be supported by a stand-alone microprocessor/CRT (graphics) combination. Some of the output from this system are presented in Figure 7.

Pr. : 2011010111 - 012600 - 025101

7-304, 2126 I. 501

```

FILE C IF YOU WANT TO CONTINUE RUN CONTINGENCY TEST OFF
FILE C IF YOU DO NOT WANT TO CALCULATE RUN CONTINGENCY
FILE C IF YOU WANT CONTINGENCY ON CONTINGENCY DATA, SP OUT FILE = C
AT THE END RUN ON DATA FILE = 25

```

Page 1 of 1

1- 7 1,6,2,5,0,0, 00 E 2 S

S TIME1 TO TIME1 TIME2 TO TIME2 LEFT LEFT: TOP0 TAIL WERE ALL SO IT DING OUT
THE

20100	01	04517	00611	13	02438	426	256	222	0	215	20	0	57	233
20100	02	05112	00705	34	06014	426	222	255	0	242	20	7	51	300
20100	03	06102	00855	09	11329	445	157	241	15	211	20	12	45	300
20100	04	07107	01015	21	11439	414	157	222	20	206	20	17	37	300
20100	05	08105	01050	04	11509	414	215	224	15	224	20	15	25	300
20100	06	09105	01020	24	11479	356	142	231	15	225	20	22	40	200
20100	07	10110	01022	22	12525	415	159	224	20	224	20	0	40	254
20100	08	11106	01035	15	12021	445	142	225	15	213	20	32	40	300
20100	09	12105	01025	13	12079	415	157	227	15	217	20	33	15	200
20100	10	01110	01040	12	12229	417	212	242	20	219	20	11	40	300
20100	11	10106	01045	29	12129	444	141	221	15	211	20	24	35	300
20100	12	10105	01120	25	12479	509	141	251	15	235	20	2	40	300
20100	13	10102	01100	35	12279	444	142	225	15	215	20	4	35	300
20100	14	10104	01150	35	12579	427	142	204	15	204	20	0	25	300
20100	15	10105	01107	26	11859	421	112	244	10	219	20	115	35	300
20100	16	10105	01107	30	11859	421	112	244	10	211	20	22	20	229
20100	17	11105	01107	30	11859	444	242	215	15	227	20	2	35	215
20100	18	11105	01107	30	11859	411	300	255	10	241	20	4	25	200
20100	19	11105	01105	14	12379	353	212	207	10	240	20	53	27	200
20100	20	10105	01105	14	12379	407	215	217	15	207	20	43	15	300
20100	21	10105	01105	14	12379	407	215	217	15	207	20	43	15	300
20100	22	10105	01105	14	12379	407	215	217	15	207	20	43	15	300
20100	23	10105	01105	14	12379	407	215	217	15	207	20	43	15	300
20100	24	10105	01112	40	12479	407	242	201	15	216	20	0	52	202
20100	25	10105	01127	02	13079	415	300	252	20	253	20	0	24	222
20100	26	10105	01127	42	13279	427	215	252	15	227	20	2	50	200
TOTAL					11302	19125	16547	737	20171					
TOTAL					5329	745	290	1633						

1.2 T T.2, 2.5, 0.0, 0.0 0.0 0.0 0.0

22 - 24th EPR

22-00000

Figure 7 - Sample CRT Display of Rough-Cut Driver Schedule Produced by Z-80 Microcomputer

REFERENCES

- [1] Bodin, L.D., and Rosenfield, D., Estimation of the Operating Cost of Mass Transit Systems, Report No. WAHCUPS--UMTA-1-76, U.S. Dept. of Transportation, Washington, D.C., September 1976.
- [2] Jenkins, R.T., "An Automated Technique for Scheduling Motormen and Conductors for the New York City Subways," OSRA/TSS Workshop on Automated Techniques for Scheduling of Vehicle Operators for Urban Public Transportation Services, Chicago, April 1975.
- [3] Orloff, C., "Route Constrained Fleet Scheduling," Transportation Science, Vol. 10, No. 2, pp. 149-168, May 1976.
- [4] Eliar, Samy E.G., A Mathematical Model for Optimizing the Assignment of Man and Machine in Public Transit "Run-Cutting", West Virginia University Engineering Experiment Station Research Bulletin No. 81, September 1966.
- [5] Guha, D., and Browne, I., "Optimal Scheduling of Tours and Days Off," ORSA/TSS Workshop on Automated Techniques for Scheduling of Vehicle Operators for Urban Public Transportation Services, Chicago, April 1975.
- [6] Rubin, J., A Technique for the Solution of Massive Set Covering Problems with Application to Airline Crew Scheduling, IBM Philadelphia Scientific Center Report 320-3004, September 1971.
- [7] Wilhelm, E., Overview of the RUCUS Package Driver Run-Cutting Program-Runs, MTR-6803, the MITRE Corporation, McLean, Virginia, December 1974.
- [8] Bennett, Brian T., and Potts, Renfrey B., "Rotating Roster for a Transit System," Transportation Science, Volume 2, No. 1, pp. 14-34, February 1968.
- [9] Bergmann, Dietrich R., Issues in Urban Bus Systems Schedule Optimization Report to Michigan Bureau of Transportation, NTIS Report No. PB 214-473, May 1972.

- [10] Levin, A., "Scheduling and Fleet Routing Models for Transportation Systems," Transp. Sci., 5, 232-255 (1971)
- [11] Bennington, G., "An Efficient Minimal Cost Flow Algorithm," Man. Sci., 19, 1042-1051 (1973).

AN APPLICATION OF BRANCH AND BOUND METHOD
TO OPTIMIZE INTERDEPENDENT PUBLIC TRANSIT NETWORK

Inwon Lee

Korea Institute of Science & Technology
Regional Development Research Center
P.O.Box 131 Dong Dae Mun
Seoul, Korea

ABSTRACT. In recent years, significant advances have been made enabling travel demand analysis and network design methods to be used as increasingly realistic evaluation tools. What has been lacking is the integration of travel demand analysis with network design models.

This paper presents an integrated(advanced) modelling system that can be used to design public transit network. Instead of assuming that trips from a zone to a workplace are fixed, the travelers' free choice of mode and destination is introduced into an optimal searching procedure using Branch and Bound method.

In order to provide an empirical test result, a joint choice probabilistic model is developed using 1970 census data. The impacts of public transit network changes on the spatial distribution of activities is measured simultaneously with the impacts on mode choice behavior. Next, by embedding the joint choice model into a discrete network optimization procedure, the best possible network investment is searched in a forward seeking procedure.

The most significant result of this study lies in the use of a linear approximation technique to improve the efficiency of Branch and Bound method. The necessary conditions of using the proposed linear approximation technique to find the globally optimal solution is mathematically demonstrated and the significance of efficiency improvements is empirically investigated. The purpose of this paper is to demonstrate that the improved capability of the discrete network optimization method makes it practically feasible to find the best network investment program based on non-linear behavioral models.

1. INTRODUCTION

In urban transportation planning, considerable effort has been concentrated on the development and application of mathematical models. Even though the models still have many limitations, the need for improved tools for transport network investment decisions is growing rapidly in the public sector.

As an example, the Regional Transportation Authority (RTA) of Northeastern Illinois has identified approximately 120 rapid transit alternatives that can be introduced to the transit network as new routes or as replacements for existing lines. Since it is not politically feasible to invest capital in two lines serving the same area, a decision was made to group the 120 alternatives into 30 major corridor alternatives with one candidate alternative being selected for each corridor.

The task of RTA capital investment planning is to select a subset from the 30 project universe which represents the best use of the RTA capital budget. Best is determined according to several objectives; maximization of transit patronage, consumers' surplus, regional accessibility, and minimization of subsidy requirement.

Since projects can either compete for passengers, cooperate in carrying passengers, or do both at the same time, project interrelationships should be reflected in the project evaluation. In order to get the best result from investment planning analysis, all possible combinations of projects should be considered. However, if one were to evaluate all possible combinatorial networks that result from 30 alternatives, over 1.073 billion (2^{30}) evaluations would be required for each objective. No conventional economic analysis technique would be useful given the available RTA budget, nor would independent evaluation of each corridor alternative correctly represent the interdependent benefits of new or replacement transit alternatives. Therefore, when it is necessary to consider the effects of network interdependence upon evaluation criteria, the problem might best be analyzed in terms of a mathematical modelling approach.

The mathematical modelling approach is often applied in three different ways: (1) the "forward-seeking" method; (2) the "backward-seeking" method and (3) the so-called "adaptive planning" method. The forward-seeking method is an inductive searching approach where the future state is

exogenously predicted and an optimal network plan is developed to best serve this given future state. The backward-seeking method is a deductive searching approach where the transportation network is designed to guide regional development so that desired objectives are attained. The adaptive planning approach is a way of combining features of the above two views in order to continuously guide a region's transportation system and its environment. The advantage of this adaptive process is in achieving a better understanding of the time dimension or future uncertainty and using this to improve effectiveness in modifying reality. The concept of optimizing a discrete network system under various alternative futures and implementing only the most robust alternative seems to be one of the most promising applications of this adaptive planning approach. However, this approach generally requires designing and implementing a very efficient and extensive mathematical modelling system.

The current status of transportation modelling analysis does not meet the standards of an extensive yet efficient modelling system. The sequential modelling process of trip generation, trip distribution, mode choice and link assignment is an expensive, error propagating procedure which inconsistently utilizes the transportation level of service variables. In many cases, the trip generation and distribution procedures are insensitive to public transit network changes. Thus many standard transport network optimization models often overlook the need for travel behavior analysis. Harris raised the question of whether these mathematical models truly represent the problem whose solution is being attempted (Harris, 1970). The conventional network design models based on the normative link assignment and a fixed O-D trip table have been criticized for use in transit network design because passengers are not all captive nor do they behave so as to achieve the system optimal point. Therefore, it is necessary to apply a descriptive link assignment (or behavioral link choice) procedure with variable O-D trip interchanges when optimizing public transit networks.

The discrete network optimization approach, such as the Branch and Bound method, can be utilized for the analysis of travel demand behavior (Steenbrink, 1974; Leblanc, 1975; Boyce, 1977). However, most branch and bound tree searching techniques are too inefficient to be used outside of the research phase because the computation time required for defining the global optimal solution grows exponentially as the number of alternatives increases. Therefore, heuristic

rules which do not find exact optima are popular in many practical network design studies. If the global optimal solution cannot be found by heuristic rules, it becomes necessary to formalize the tradeoff between the cost of computation and the closeness to the global optimum. However, the formalization which is necessary to decide when to stop searching has rarely been studied based on real world network design problems.

Besides the above arguments, there are many other questions or problems in the use of the mathematical modelling approach. Data availability, forecasting errors, grossness, complicatedness, and mechanicalness are typically listed among others (Lee, 1973). The purpose of this paper is to examine empirically the performance of an improved mathematical modelling approach and to provide insights for the future development of mathematical models.

The first part of this paper focuses on the development of a behavioral demand model and an efficient transit network optimization method which correctly represents travel demand changes. The second part of the paper investigates empirically the merits and shortcomings of the use of a discrete optimization approach in conjunction with the travel behavior models to search inductively for the optimal investment strategy in improving the existing public transit network. The overall evaluation of the mathematical modelling approach developed might be based on the preferences and goals of decision makers. However, the performances of the models developed herein are reported in quantitative terms.

2. DEVELOPMENT OF A JOINT CHOICE MODEL FOR TRANSPORTATION PLANNING

2.1. Model Structure

The evaluation and selection of alternative public transit network plans are generally dependent on the prediction of their economic, environmental and social impacts. In the past, information on the expected impacts of proposed plans was largely based on judgmental information provided by various experts or government planners. The process of reaching consensus (or agreed-upon impact statements), however, is too unwieldy to use for the evaluation of a large number of alternative networks. In addition, the judgmental evaluations were often focused on the first order direct impacts. Therefore, the need for a quantitative modelling approach has long been recognized not only to explore mutually interdependent indirect (or secondary) impacts but also to provide estimates of the impacts of changes in transport network systems.

At the present time, the standard transportation modelling process (i.e., the sequential modelling process of trip generation, trip distribution, mode choice and link assignment) for network impact analysis is associated with two major shortcomings: the expensive and error propagating nature of its utilization, and the question of its theoretical validity. Even though the propagation of errors from one sub-model to another may be a minor issue, the high cost of applying the entire demand modelling process (such as UTPS Package) greatly decreases its applicability in transport network design (or optimization) studies because it is likely to be used only a limited number of times. Theoretical questions concerning the validity of the demand estimate process have been discussed in numerous papers (Stopher, Lisco 1970; Brand, 1972; Robert, 1972; Lee, 1973; Ben-Akiva, 1973; McFadden, 1973; Hutchinson, 1974; Lerman, 1975; Cesario, 1977 for example). The inconsistency of using major policy variables such as travel time and cost, the insensitivity of models to level of service changes, and the improper treatment of the conditional probabilities have been most widely criticized.

To derive a joint(destination and mode) choice model the following entropy maximization approach is proposed:

$$\text{Maximize } E = \frac{T!}{\prod_{i,j,k} T_{ijk}!} \quad (1)$$

$$\text{Subject to } \sum_i \sum_j \sum_k T_{ijk} U_{ijk} = U \quad (2)$$

$$\sum_i \sum_k T_{ijk} = EMP_j \quad (3)$$

Where T is the total number of trips.

T_{ijk} is the trip interchange between zone i and j by mode k .

U_{ijk} is the per trip utility from i to j by mode k .

U is total system utility.

EMP_j is total employment in zone j .

If we maximize the objective function by taking logarithms (i.e., $\ln(T!) - \sum_i \sum_j \sum_k \ln(T_{ijk}!)$) subject to (2) and (3), the Lagrangian maximization problem based on Stirling's approximation becomes:

$$\begin{aligned} \ln(E) = & - \sum_i \sum_j \sum_k T_{ijk} \ln(T_{ijk}) + \sum_j \lambda_j [EMP_j - \sum_i \sum_k T_{ijk}] \\ & + \beta [U - \sum_i \sum_j \sum_k T_{ijk} U_{ijk}]. \end{aligned} \quad (4)$$

by taking the partial derivative $\frac{\partial \ln(E)}{\partial T_{ijk}}$ and setting it equal to zero we obtain:

$$T_{ijk} = \exp(-\lambda_j) \exp(-\beta U_{ijk}) \quad (5)$$

Where $\exp(-\lambda_j) \equiv EMP_j / \sum_{ik} \exp(-\beta U_{ijk})$ from the constraint

equation (3). This relationship is needed to satisfy (3).

λ_j, β are Lagrangian multipliers.

Therefore, the simultaneous (or joint) place-of-home and mode choice model can be shown as follows:

$$T_{ijk} = EMP_j \frac{\exp(-\beta U_{ijk})}{\sum_i \sum_k \exp(-\beta U_{ijk})} \quad (6)$$

As Wilson argued in his singly constrained model Wilson, 1974, we may assign zonal weight based on zonal attractiveness or simply number of residents in zone 1 to the utility term in both numerator and denominator. If this zonal weight is introduced based on zonal attractiveness, the equation (6) will be changed as follows:

$$T_{ijk} = EMP_j \frac{A_i \exp(-\beta U_{ijk})}{\sum_i \sum_k A_i \exp(-\beta U_{ijk})} \quad (7)$$

Where A_i is the attractiveness of zone 1

This deviation from the equation (6) can be interpreted as a redefinition of the utility function. Since $A_i = \exp(\ln(A_i))$, the weighted utility term in equation (7), (i.e. $A_i \exp(-\beta U_{ijk})$), is identical to $\exp(\ln(A_i) - \beta U_{ijk})$. In other words, if the zonal attractive measures are introduced in the constraint equation (2), we can obtain equation (7). The further detailed discussion of the utility function and independent variables are included in the next section.

The probabilistic interpretation of the joint destination and mode choice model defined as equation (7) is:

$$\text{Prob}(i,j,k) = \frac{A_i \exp(-\beta U_{ijk})}{\sum_i \sum_k A_i \exp(-\beta U_{ijk})} \quad \text{for every } j.$$

$\text{Prob}(i,j,k)$ can be directly utilized for the journey from-work-to-home matrix if the parameters of the utility function are determined. It is now obvious that this joint choice probability of trip distribution between i and j via mode k is not fixed but changes in a non-linear fashion depending upon the utility function or zonal attractiveness changes. This is one of the most desirable characteristics of the model in that work trip interchange and mode choice are defined as a direct function of mode specific utilities.

If a Cobb-Danglas utility function is introduced instead of U_{ijk} , the proposed model structure would result in the following form:

$$T_{ijk} = EMP_j \frac{P_i \exp[\alpha Y_i + \beta t_{ijk} + \gamma c_{ijk} + \delta D_{iCBD} + a_1 X_1 + a_2 X_2]}{\sum_n \sum_m P_n \exp[\alpha Y_n + \beta t_{njm} + \gamma c_{njm} + \delta D_{nCBD} + a_1 X_1 + a_2 X_2]}$$

Where T_{ijk} is the trip interchanges between i and j via mode k.

EMP_j is employment in zone j.

P_i is the number of residents of zone i.

Y_i is the average household income of zone i.

t_{ijk} is average door-to-door travel time from i to j via mode k.

c_{ijk} is average door-to-door cost of travel from i to j by mode k.

D_{iCBD} is air-line distance from zone i to CBD.

X_1, X_2 are mode specific dummy variables. If $X_1 = 1$, auto is used and if $X_2 = 1$, commuter rail is used, if both are zero then rail transit (CTA) is used.

$\alpha, \beta, \gamma, \delta$ are coefficients to be obtained through calibration.

2.2 Calibration Methodology

As discussed in numerous papers, there are two basic calibration methods for determining the parameters of the utility function: minimize variance (least squares) or maximize likelihood. The least squares method minimizes

$\sum_{ik} [T_{ijk} - \bar{T}_{ijk}^0]^2$, \bar{T}_{ijk}^0 is the observed trip interchanges between i and j via mode k while the maximum likelihood method maximizes

$$\prod_{ik} \left(\frac{T_{ijk}}{T_j} \right)^{T_{ijk}} \text{ where } T_j \text{ is total}$$

observed trip for employment zone j.

The most efficient method for calibrating the non-linear model involves calculating partial derivatives for all the coefficients (or parameters), and simultaneously solving all the partial derivative equations, after setting these equal to zero. However, the analytical method for solving such a set of simultaneous equations is extremely difficult and has not yet been successful.

Alternatively, iterative searching procedures such as Newton-Raphson, Davidon-Fletcher-Powell, or Fletcher-Reeves method [Fletcher and Reeves, 1964] have been widely utilized. These techniques basically search a surface to determine which coefficient values would either minimize variance or maximize likelihood. In detail, they operate as follows: (a) estimate parameter values; (b) calculate partial derivatives for the parameters; (c) determine, based on the largest magnitude derivative (positive or negative) which coefficient value should be changed; (d) determine the change in the parameter based on the move which would bring the derivative value closest to zero; (e) repeat the above steps until the stopping criteria (which is the closeness of the partial derivatives to zero) is satisfied.

For this research to calibrate the joint destination and mode choice models, the Fletcher-Reeves method is selected. The Fletcher-Reeves method, which does not require the inversion of the Hessian matrix for each searching step, was found flexible enough (and also suitable) for handling the mode specific constants independently from other joint choice variables. However, it is wrong to assert that other methods are inferior to the Fletcher-Reeves method. It should be noted, however, that the Fletcher-Reeves gradient method utilizing the first partial derivative equations is much more computer efficient than the quasi-Newton method which approximates the first partial derivatives.

2.3 Data Collection

Considering the fact that our purpose in demand model development is to apply it for network optimization, where zonal aggregation is necessary, the modelling of the behavior of groups instead of individual behavior is considered more acceptable and practical for this research especially since adequate data on individual travelers was not available. The conventional approach defining groups' behavior in probabilistic term based on a one square mile zone system is considered as a proper replica of trip distribution and mode

choice patterns.

For the Northeastern Illinois Region, a 25 percent random sample of the CATS zone system was drawn. This sample contains 168 zones. With the exception of two zones in the fringe of the study area, the selected zones are about one square mile zones.

The base data associated with the proposed sample set were both collected from existing sources and developed from engineering measurements. The data collected from existing sources are zonal population, zone income, CBD employment, and trip interchanges by mode from the samples zone to the CBD. The travel time and cost by the different modes from the zone centroid to the CBD were measured based on a simple minimum path procedure.

The zonal population was the 1970 residential zone population as reported in the census. The zone income represents the average family income in the zone for 1970 as reported in the census. As with population, the census income data were converted into the CATS zone scheme.

The CBD employment data was based on the work trip data, and was calculated by summing the number of work trips by all modes from the sample zones to the CBD. The zonal trip interchange data was based on the special section of 1970 census, i.e. Urban Transportation Planning Package (UTPP) which contains tabulations of the 1970 journey-to-work data.

To estimate the utilities of the four line-haul modes i.e. auto, bus, rapid transit and commuter rail, various existing data sources and specially developed methodologies are utilized. Theoretically the utilities of each mode include a range of factors, namely its time, cost, reliability, comfort, safety and convenience. For the demand model calibration, the time and cost values were estimated directly, while the other factors were initially indirectly accounted for through the use of distance variables and mode specific constants.

2.4. Calibration Results

Upon the completion of the required data collections and modifications, a series of joint destination (residential zone) and mode choice models are calibrated in this section. In order to determine the exact model parameters, utility function variables, and calibration criteria, five alternative joint choice model formulations, each with a slightly different underlying rationale have been specified and examined

based on the Chicago CBD work trip data. From these five alternative models, the final joint choice model was chosen for the next part of study, i.e., the optimization of a public transit network with a travel demand model.

Table 2.1 shows the parameter values for each model structure. Three general comments concerning the models can be made: (a) the signs of time, cost and income variables are correct; (b) r^2 s for the models are approximately equal and all quite good; and (c) the mode specific constants show that there are biases in the measures of other variables.

Based on these results in Table 2.1, a two step process was followed to determine which model (i.e. independent variables) should be used. The first step involved a comparison between the models with and without income. The following points were noted for those models which include income: (a) there is a slightly higher r^2 ; (b) the constant terms are significantly larger and thus have a greater impact on mode choice (conversely, these models are less sensitive to the level of service variables); and (c) the value of time is much smaller and is thought to be inconsistent with other researches in the Chicago region [Lisco, 1967; Wigner, 1973]. In addition, to be able to use the models with income for predicting travel demand in the future it would be necessary to forecast household income. Thus it was decided to use the models which do not contain an income variable in the next part of the research.

The second step was a comparison of Models One and Three, neither of which have income variables. Model three, which includes mode specific constants was selected for the following reasons: (a) r^2 is slightly larger for Model Three; (b) the consistency of the parameter values for the different calibration methods is greater for Model Three, and (c) it seems that measured utilities do not sufficiently represent actual utilities and thus the concept of adjustment to measure times and costs through mode specific constants is necessary.

Therefore, the utility function selected for this research is

$$U_{iCBDk} = [a_k X_k - \beta \ln(t_{iCBDk}) - \gamma \ln(c_{iCBDk})].$$

TABLE 2.1 CALIBRATION OF SIMULTANEOUS MODELS

MODEL	CALIBRATION METHOD	TIME		COST		CONSTANT			INCOME			R ² or MLI
						Auto	Commuter Rail	Transit	Auto	Commuter Rail	Transit	
ONE	Least Squares	-.864	-.595	-	-	-	-	-	-	-	-	R ² = .74
	Maximum Likelihood	-.946	-.363	-	-	-	-	-	-	-	-	MLI = .9684 (R ² = .71)
TWO	Least Squares	-.809	-.666	-	-	-	-	-	.337	.337	.337	R ² = .78
	Maximum Likelihood	-.984	-.369	-	-	-	-	-	.331	.331	.331	MLI = .9707 (R ² = .72)
THREE	Least Squares	-.841	-.326	-.126	-.148	.275	-	-	-	-	-	R ² = .76
	Maximum Likelihood	-.884	-.409	.134	-.206	.072	-	-	-	-	-	MLI = .9695 (R ² = .73)
FOUR	Least Squares	-.639	-.696	.237	-.373	.136	.405	.405	.405	.405	.405	R ² = .80
	Maximum Likelihood	-.857	-.797	.431	-.426	-.004	.398	.398	.398	.398	.398	MLI = .9726 (R ² = .76)
FIVE	Least Squares	-.599	-.882	.766	-.644	-.123	.117	.520	.502	.502	.502	R ² = .81
	Maximum Likelihood	-.827	-.532	.343	-.540	.197	.318	.510	.210	.210	.210	MLI = .9729 (R ² = .74)

This would be the simplest possible utility function that can be utilized in the public transit network design study.

The final analysis involved the determination of which calibration method should be used. As was shown in the Table 2.1, r^2 is always higher for the least squares method, while the consistency of the time and cost parameter values is substantially greater for the maximum likelihood calibration. These two facts can be interpreted to mean that the least squares method better replicates existing data, but that its parameter values are inconsistent (or statistically inefficient and biased).

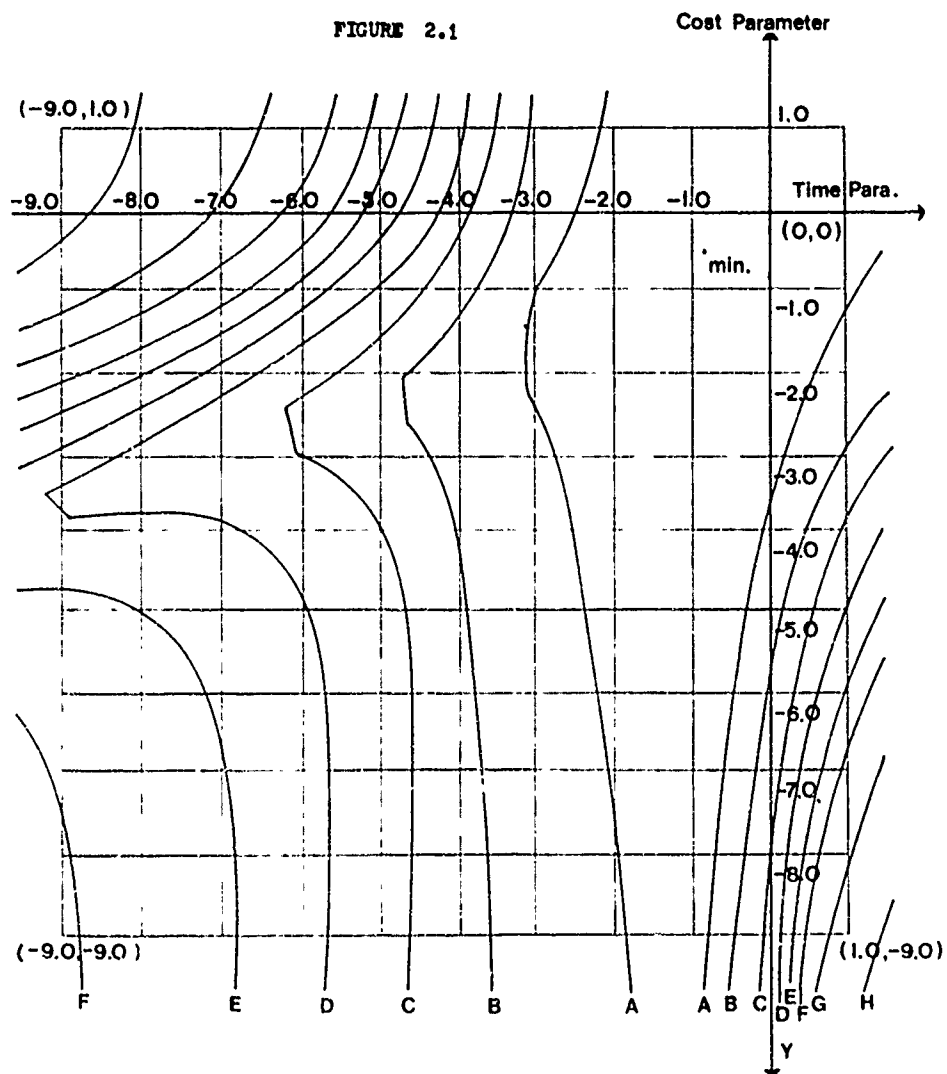
This statistical bias is important in that the smaller the variation in the parameter values, the greater the reliability of any one specific parameter value. Examination of the coefficient of the most important variable, travel time, shows that for the maximum likelihood method, the variation is from $-.827$ to $-.904$ while for least squares estimation, the variation is from $-.599$ to $-.864$.

A related point is the sensitivity of the different calibration criteria to changes in parameter values. More sensitive measures are preferred since they imply a smaller range within which the final coefficients could lie. Figures 2.1 and 2.2 illustrate the sensitivity of both calibration criteria. In the figure, each contour line represents a 10 percent decrease in the unexplained variance, or a 10 percent increase in the value of the likelihood function. For least squares calibration (Figure 2.1), the total variance as well as the amount of explained variance associated with different calibration coefficients are calculated, and the percentiles which represent the amount of unexplained variance are drawn. For maximum likelihood calibration (Figure 2.2), the total likelihood indicators associated with different calibration coefficients are calculated, and percentiles which present the amount of likelihood are drawn.

The figures show that the maximum likelihood estimate is far more sensitive to parameter values than is the least squares criterion. In addition, the plane associated with the maximum likelihood method is close to a normal distribution, while the least squares plane has an unknown skewed distribution. This documents the belief that the maximum likelihood method provides less statistically biased parameters than does the least squares method.

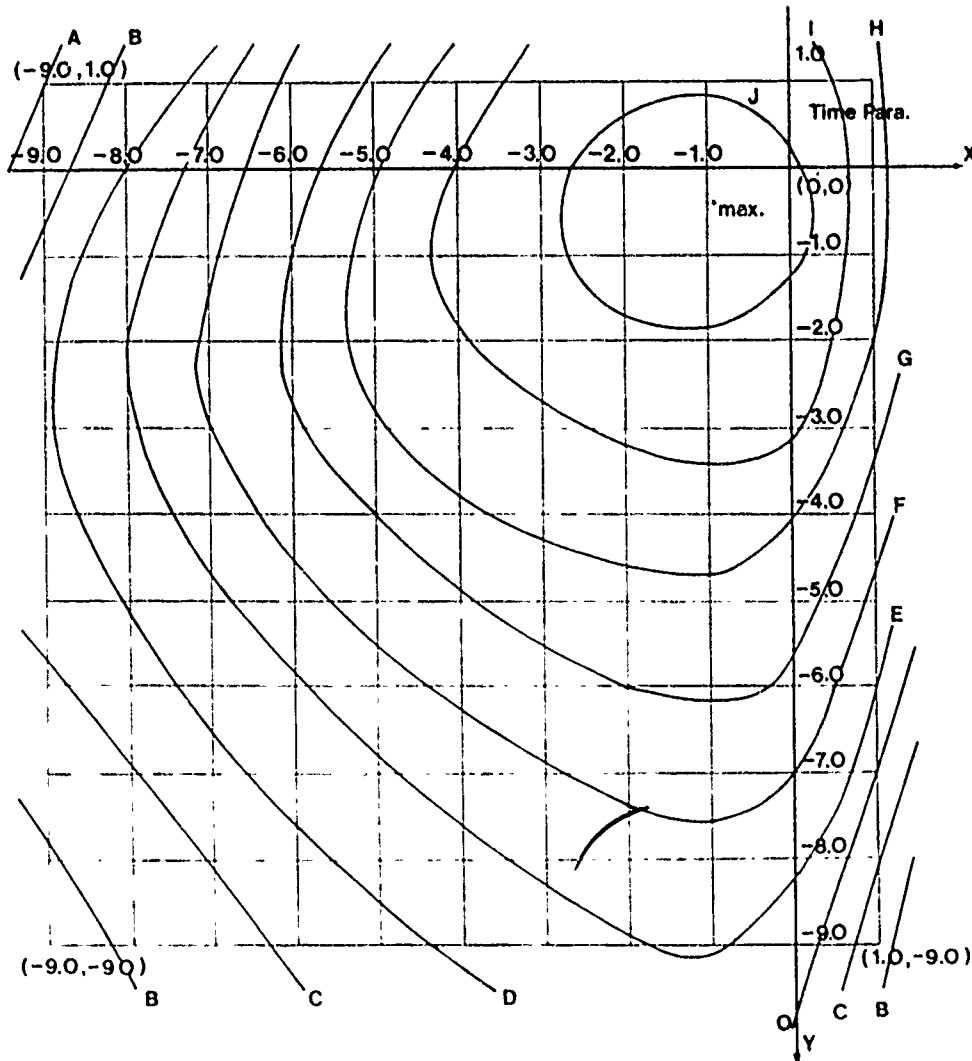
Based on the above analysis it was determined that the maximum likelihood calibration should be used for the public

FIGURE 2.1



Least Square Estimation Contours

FIGURE 2.2



Maximum Likelihood Estimation Contours

3. EMBEDDING THE TRAVEL DEMAND MODEL INTO NETWORK OPTIMIZATION PROCEDURES

3.1. Overview of Network Optimization Methodology

The development of an optimal regional network system has associated with it several major conceptual difficulties which must be addressed in the formulation of an optimization model. Generally the problems involved with transport network optimization can be grouped into two major categories:

a. Quantification of Optimization Criteria

The difficulties in applying mathematical programming techniques to social problems involve the definition of the objective function and constraints, the measurements of parameters, and the tendency to oversimplify problems. As Harris points out (1970), the real obstacle is that there is no guarantee that the problem as formulated according to any particular programming method is a true representation of (isomorphic with) the problem whose solution is being attempted. Ignorance about travel behavior and some important decision criteria in optimization are two main sources causing the lack of this isomorphism. The minimization of total tripmakers' travel time under a fixed trip interchange table is a typical approach which oversimplifies the transport network design problem. Patronage, consumers' surplus, natural resource savings and environmental concerns are also important perspectives that should be considered in practical transport network optimization.

b. Interdependence among Alternatives

Given that a criterion for benefit is selected (consumer's surplus or patronage for example), the computed benefit of an alternative will vary depending on whether other proposed alternatives are assumed to be in existence. The assumption that the linear sum of each individual link benefit is the benefit of the entire network calculated in a certain way, is incorrect in most cases. The use of independently quantified link benefits would lead us to myopic decision making. A solution to the problem of interdependence among links would involve the direct evaluation of all the combinatorial network possibilities. The number of the feasible networks is very large, ranging upward from about $\binom{n}{k}$ where n is the number of alternatives proposed and k is the number of alternatives to be selected. If we can evaluate one alternative network per second under unusually favorable circumstances, to select 15 alternatives out of 30

transit network optimization study. The final model to be used is therefore:

$$T_{iCBDk} =$$

$$EMP_{CBD} \frac{P_i(t_{iCBDk})^{-.884} (c_{iCBDk})^{-.409} \exp[.134X_1 - .206X_2 + .072X_3]}{\sum_i \sum_k P_i(t_{iCBDk})^{-.884} (c_{iCBDk})^{-.409} \exp[.134X_1 - .206X_2 + .072X_3]}$$

will take about eight and a half years. Therefore an improvement in discrete network optimization techniques should be made for properly defined network design problems.

Because of these difficulties, the selection of a transit network optimization technique necessitates compromises between the goal of providing a realistic framework upon which to base evaluation but limiting the approach to one that is economical and computationally efficient.

There have been many approaches developed to optimize the interdependent(or non-linear) network problem. Among others the branch and bound or the backtrack programming algorithm have been most widely applied in various practical network optimization problems(Ridley, 1965; Ochoa-Rosse, 1968; Scott, 1969; Boyce, 1973; Steenbrink, 1974; Leblanc, 1975)

The Ochoa-Rosse method was developed to minimize total users' travel time subject to a capital investment(budget) constraint. Although the total users' travel time is calculated based on a fixed trip interchange table and is assumed to be monotonically decreasing when more links are added to the base network(i.e., nonexistence of Braess' paradox), the users' path choice behavior can be realistically considered in their branch and bound or backtrack algorithm.

In order to further improve the efficiency of enumeration, Ochoa-Rosso proposed two approaches of ordering the operation of branching. Instead of searching the tree in a clockwise and downward direction, one approach is to branch from the active node with the lowest value for F_i and the other approach is to branch always from the last active node generated. The first approach is commonly called "bounding operation" and the second "backtracking operation". The relative efficiency of the backtracking operation was reported by Steenbrink(1974) as follows:

1. Since only the last solution and feasible solution with the least upper bound need to be stored, minimum storage space is required in computer application.
2. Since the branching scheme of the backtracking operation has been fixed in advance, the branching is not the most efficient.

In other words the backtracking operation would take more computation time but requires less storage space than the bounding operation.

However, due to the restriction that network evaluation still has to be done for every active node to set upper bounds, an alternative approach or a modification of the branch rejection operation is therefore necessary.

A method of using the link-specific(or marginal) benefit measures is proposed to modify the rejection operation in the following section.

3.2. Modification of the Implicit Enumeration Method with a Linear Approximation Technique

Instead of using the actual network evaluation for the rejection operation, the simple sum of link-specific(or marginal) benefit measures might be a useful approximation of the network benefit function $F(x)$.

If the sum of link-specific benefits, i.e., $\sum B_i X_i$, is greater than or equal to actual network benefit $F(x)$, this sum can be compared to a possible maximum value of $F(x)$. As long as $\sum B_i X_i < F(x^*)$, then $F(x) < F(x^*)$. Therefore only solutions whose linear sum of benefits are greater than the current lower bound need to be examined for further branching in the implicit enumeration method.

If the multinomial logit model and all-or-nothing assignment are accepted as simulating traveler behavior, the following lemmas allow substitution of a linear approximation of benefits for actual network evaluations. In order to develop these results, some definitions are necessary.

Definition: Links A and B are "competing" if the minimum path travel time between every O-D pair cannot be reduced by the use of both links A and B. Mathematically, $\min(t_{ij}^A, t_{ij}^B) = t_{ij}^{A,B}$ for every (i,j) . Where t_{ij}^A is the minimum path travel time from i to j when link A is added to the existing network and t_{ij}^B is the minimum path travel time from i to j when link B is added to the existing network. $t_{ij}^{A,B}$ is the minimum path travel time from i to j when link A and B are simultaneously introduced into the existing network.

Link A and B are "complementary" if $\min(t_{ij}^A, t_{ij}^B) > t_{ij}^{A,B}$, for at least one (i,j) .

Lemma 1. If a transit link A is "competing" with another transit link B, the transit patronage increase due to link A addition(i.e., $d(A)$) and the transit patronage increase due to link B addition(i.e., $d(B)$) have the following relationship:

$$d(A) + d(B) \geq d(A,B)$$

Where $d(A,B)$ is the patronage increase due to the simultaneous introduction of link A and B.

PROOF: Let $G(E)$ be the transit trips(mode k) generated from zone i under the existing network E and $G_1(A)$ and $G_1(B)$ be the transit trips generated from zone i when link A and link B are respectively introduced into the existing system. Then based on the joint destination and mode choice model, $G_1(E)$, $G_1(A)$ and $G_1(B)$ are defined as follows:

$$G_i(E) = \sum_j T_{ijk}^E = \sum_j \left(EM_j P_i \exp(\alpha \ln t_{ijk}^E) / \sum_m \sum_n [P_n \exp(\alpha \ln t_{njm}^E)] \right) \quad (1)$$

$$G_i(A) = \sum_j T_{ijk}^A = \sum_j \left(EM_j P_i \exp(\alpha \ln t_{ijk}^A) / \sum_m \sum_n [P_n \exp(\alpha \ln t_{njm}^A)] \right) \quad (2)$$

$$G_i(B) = \sum_j T_{ijk}^B = \sum_j \left(EM_j P_i \exp(\alpha \ln t_{ijk}^B) / \sum_m \sum_n [P_n \exp(\alpha \ln t_{njm}^B)] \right) \quad (3)$$

Where T_{ijk}^E , T_{ijk}^A , and T_{ijk}^B are the total trips going from zone i to zone j by mode k under the cases E , A and B respectively.

P_i is the attractiveness of zone i measured by zonal population
 EM_j is the total number of employment of zone j .

The transit trips generated from zone i when link A and link B are simultaneously introduced to the existing system, i.e., $G_1(A,B)$, can also be defined as follows:

$$G_i(A,B) = \sum_j T_{ijk}^{A,B} = \sum_j \left(EM_j P_i \exp(\alpha \ln t_{ijk}^{A,B}) / \sum_m \sum_n [P_n \exp(\alpha \ln t_{njm}^{A,B})] \right) \quad (4)$$

Since link A and link B are competing, $\min[t_{ijk}^A, t_{ijk}^B] = t_{ijk}^{A,B}$ or $\exp(\alpha \ln t_{ijk}^{A,B}) = \max[\exp(\alpha \ln t_{ijk}^A), \exp(\alpha \ln t_{ijk}^B)]$ since α is negative. However, $\sum_m \sum_n \exp(\alpha \ln t_{njm}^{A,B}) \geq \max[\sum_m \sum_n \exp(\alpha \ln t_{njm}^A), \sum_m \sum_n \exp(\alpha \ln t_{njm}^B)]$ because possibly $\exp(\alpha \ln t_{njm}^{A,B}) > \exp(\alpha \ln t_{njm}^A)$ and $\exp(\alpha \ln t_{njm}^{A,B}) > \exp(\alpha \ln t_{njm}^B)$ for at least one $n \neq i$ or $m \neq k$.

It is clear that when each zone is assigned by the "all-or-nothing" method to the minimum time path link that the market area of link A, i.e., $M(A)$ and that of link B, have the following relationship:

$$M(A) \cup M(B) = M(A, B).$$

If $n \in M(B)$, then $t_{njm}^A > t_{njm}^{A,B}$ or $\exp(\alpha \ln t_{njm}^{A,B}) > \exp(\alpha \ln t_{njm}^A)$.

Therefore,

$$\begin{aligned} G_i(A, B) &= \sum_j \left(EM_j P_i \exp(\alpha \ln t_{ijk}^{A,B}) / \sum_m \sum_n [P_n \exp(\alpha \ln t_{njm}^{A,B})] \right) \\ &\leq \sum_j \frac{EM_j P_i \text{MAX}[\exp(\alpha \ln t_{ijk}^A), \exp(\alpha \ln t_{ijk}^B)]}{\text{MAX}[\sum_m \sum_n P_n \exp(\alpha \ln t_{njm}^A), \sum_m \sum_n P_n \exp(\alpha \ln t_{njm}^B)]} \\ &= \sum_j EM_j P_i \text{MAX} \left[\frac{\exp(\alpha \ln t_{ijk}^A)}{\text{MAX}[\sum_m \sum_n P_n \exp(\alpha \ln t_{njm}^A), \sum_m \sum_n P_n \exp(\alpha \ln t_{njm}^B)]}, \right. \\ &\quad \left. \frac{\exp(\alpha \ln t_{ijk}^B)}{\text{MAX}[\sum_m \sum_n P_n \exp(\alpha \ln t_{njm}^A), \sum_m \sum_n P_n \exp(\alpha \ln t_{njm}^B)]} \right] \\ &\leq \sum_j EM_j P_i \text{MAX} \left[\frac{\exp(\alpha \ln t_{ijk}^A)}{\sum_m \sum_n P_n \exp(\alpha \ln t_{njm}^A)}, \frac{\exp(\alpha \ln t_{ijk}^B)}{\sum_m \sum_n P_n \exp(\alpha \ln t_{njm}^B)} \right] \\ &= \sum_j \text{MAX}[T_{ijk}^A, T_{ijk}^B] \leq \sum_j (T_{ijk}^A + T_{ijk}^B) = G_i(A) + G_i(B) \quad (5) \end{aligned}$$

Summing over all zones i,

$$\sum_i G_i(A,B) = \sum_i \sum_j T_{ijk}^{A,B} \leq \sum_i \sum_j \max[T_{ijk}^A, T_{ijk}^B]. \quad (6)$$

Therefore,

$$\sum_i \sum_j [T_{ijk}^{A,B} - T_{ijk}^E] \leq \sum_i \sum_j \max[T_{ijk}^A - T_{ijk}^E, T_{ijk}^B - T_{ijk}^E]. \quad (7)$$

$$\begin{aligned} \text{Since } d(A,B) &= \sum_i \sum_j [T_{ijk}^{A,B} - T_{ijk}^E], \quad d(A) = \sum_i \sum_j [T_{ijk}^A - T_{ijk}^E] \text{ and} \\ d(B) &= \sum_i \sum_j [T_{ijk}^B - T_{ijk}^E], \quad d(A,B) \leq \sum_i \sum_j \max[T_{ijk}^A - T_{ijk}^E, T_{ijk}^B - T_{ijk}^E] \\ &\leq \sum_i \sum_j [T_{ijk}^A - T_{ijk}^E] + [T_{ijk}^B - T_{ijk}^E] = d(A) + d(B). \end{aligned} \quad (8)$$

Thus, $d(A,B) \leq d(A) + d(B)$

Q. E. D.

In the above lemma, equation 5 is a critical step because it demonstrates that $\sum_j T_{ijk}^{A,B} \leq \sum_j \max[T_{ijk}^A, T_{ijk}^B]$ for every i and k. Since trips generated in each zone will access that link with the minimum travel time, the share of trips for link A or B will be less in competition with the other line than when either exists alone. Note that according to the joint destination and mode choice model total transit trip generation will increase due to the diversions from other modes and other zones if the utility or accessibility of a zone is increased.

The above lemma and the property of the minimum time path algorithm which guarantees $t_{ijk}^{A,B} = \min(t_{ijk}^A, t_{ijk}^B)$ when links A and B are competing, lead to another important lemma as follows;

Lemma 2. If a transit link A and another transit link B are competing to serve the passengers generated from zone i, the net consumers' surplus changes of zone i residents achieved by simultaneously introducing link A and link B, i.e., $CS_i(A,B)$ has the following relationship to the changes in consumers' surplus due to introduction of individual links:

$$(1) \quad CS_i(A,B) \leq CS_i(A) + CS_i(B) \quad \text{and hence}$$

$$(2) \sum_i CS_i(A,B) \leq \sum_i CS_i(A) + \sum_i CS_i(B)$$

Where $CS_i(A)$ is the net consumers' surplus change of residents when link A is added to the existing network.

$CS_i(B)$ is the net consumers' surplus change of zone i residents when link B is added to the existing network.

PROOF: By definition of consumers' surplus,

$$CS_i(A) = \frac{1}{2} \sum_j (T_{ijk}^A + T_{ijk}^E) (t_{ijk}^E - t_{ijk}^A),$$

$$CS_i(B) = \frac{1}{2} \sum_j (T_{ijk}^B + T_{ijk}^E) (t_{ijk}^E - t_{ijk}^B) \text{ and}$$

$$CS_i(A,B) = \frac{1}{2} \sum_j (T_{ijk}^{A,B} + T_{ijk}^E) (t_{ijk}^E - t_{ijk}^{A,B}).$$

From the travel demand model, if $t_{ijk}^A = t_{ijk}^{A,B}$, then $T_{ijk}(A) \geq T_{ijk}(A,B)$.

Therefore $T_{ijk}(A,B) \leq T_{ijk}(A)$ if $\text{MAX}[(t_{ijk}^E - t_{ijk}^A), (t_{ijk}^E - t_{ijk}^B)] = (t_{ijk}^E - t_{ijk}^A)$

and $T_{ijk}(A,B) \leq T_{ijk}(B)$ if $\text{MAX}[(t_{ijk}^E - t_{ijk}^A), (t_{ijk}^E - t_{ijk}^B)] = (t_{ijk}^E - t_{ijk}^B)$.

Note $t_{ijk}^A = t_{ijk}^E$ if $j \in M(B)$ and $t_{ijk}^B = t_{ijk}^E$ if $j \in M(A)$.

Therefore $CS_i(A) + CS_i(B) =$

$$\begin{aligned} & \frac{1}{2} \sum_j [(T_{ijk}^A + T_{ijk}^E) (t_{ijk}^E - t_{ijk}^A)] + [(T_{ijk}^B + T_{ijk}^E) (t_{ijk}^E - t_{ijk}^B)] \\ & \cong \frac{1}{2} \sum_j (T_{ijk}^{A,B} + T_{ijk}^E) \times \text{MAX}[(t_{ijk}^E - t_{ijk}^A), (t_{ijk}^E - t_{ijk}^B)] \\ & = \frac{1}{2} \sum_j (T_{ijk}^{A,B} + T_{ijk}^E) (t_{ijk}^E - t_{ijk}^{A,B}) = CS_i(A,B). \end{aligned}$$

$$\text{And } \sum_i CS_i(A) + \sum_i CS_i(B) \cong \sum_i CS_i(A,B).$$

Q.E.D.

The above two lemmas lead to the following general theorem.

THEOREM

If the proposed transit links A,B,C,..., K are all competing with one another as per the previous definition and transit patronage and consumers' surplus are quantified marginally for each individual link by adding each link independently to the existing network, then

- (1) $d(A) + d(B) + d(C) + \dots + d(K) \geq d(A,B,C, \dots, K)$
- (2) $CS(A) + CS(B) + CS(C) + \dots + CS(K) \geq CS(A,B,C, \dots, K).$

This theorem holds as far as the transit travel time from zone i to zone j by link A,B,C,...,K,

$$(t_{ijk}^{A,B,C, \dots, K}) = \min[t_{ijk}^A, t_{ijk}^B, t_{ijk}^C, \dots, t_{ijk}^K]$$

for every i and j in the study area.

4. A CASE STUDY: THE OPTIMIZATION OF THE CHICAGO AREA RAPID TRANSIT NETWORK

4.1. Background

At the present time, a large number of rapid transit projects have been proposed for the Chicago Metropolitan Area. These projects include replacements and extensions of existing lines and the constructions of new lines. Due to the expense of constructing new lines of replacing existing facilities, not all projects can be financed under the capital budget.

After a technical feasibility study of each project and preliminary planning analyses, one potential project for each corridor was proposed. These candidate projects for major transportation corridors should be evaluated in terms of available capital budgets and total regional benefit. The benefit of investment in a corridor project cannot be measured without information concerning investment in its neighboring corridors. Interdependence among corridor projects is a major concern in the quantification of regional benefits such as transit patronage and consumer's surplus.

Therefore a regional network design method such as previously discussed must be utilized to determine the proper level of investment among corridors. The method of maximizing total transit patronage or consumers' surplus subject to an available budget is an appropriate technique for the problem. Among many questions associated with the utilization of this technique, the capability of efficiently searching for an optimal regional network under the given evaluation criteria and various alternative budgets is of prime importance.

Table 4.1 illustrates 9 selected alternatives for this study and the capital cost estimates of the 9 corridor alternatives .

Those cost estimates are derived from RTA unit costs per mile(classified by type of rail construction and difficulty of construction), unit cost per station, and rough estimates of right of way cost.

Alternative Number	Alternative Description	Capital Cost Estimates (in Millions)
1	Replacement of Existing Howard Line from North Ave. to Wilson Ave.	\$70
2	Extension of Milwaukee Line to O'Hare Airport(including subway loop at the Airport)	\$290
3	Extension/Replacement of Existing Lake Line from Halsted St. to Franklin park	\$330
4	Extension of Congress Line to Oakbrook	\$295
5	Extension of Douglas Line to Riverside	\$155
6	Extension of Ravenswood Line to Skokie	\$160
7	Extension of Englewood Line to Midway Airport	\$150
8	Proposed New Line for South- west Corridor(from Loop to Cicero)	\$230
9	Extension of Dan Ryan Line to Blue Island in the west and to Lake Calumet in the east	\$350

Table 4.1 Description of 9 Selected Alternatives

4.2. Test of Interdependence

The first issue is the importance of project benefit interdependence for optimal rail network design. The answer to this question will be more reliable if several alternative situations are considered. Therefore, as mentioned previously, two optimization criteria (i.e. transit patronage and consumer's surplus), four alternative levels of budget and two different methods of project benefit evaluation are considered, in all comprising 16 experimental cases. The following are the 16 cases and their abbreviations:

1. Case one is the maximization of total network demand with a constraint of 30 percent investment of the total maximum budget of 2.03 billion. The link-specific benefit is quantified by adding one link at a time to the existing CTA rail network system, i.e. the existing network is the base network. The abbreviation is D/30/EXT standing for Demand/30% Investment/Add to the existing network.
2. Case two is same as case one except the investment level is increased to 40 percent of 2.03 billion. The abbreviation is D/40/EXT.
3. Case three is D/50/EXT.
4. Case four is D/70/EXT.
5. Case five is to maximize consumer's surplus at 30 percent investment, as in case one. The abbreviation is CS/30/EXT.
6. Case six is CS/40/EXT.
7. Case seven is CS/50/EXT.
8. Case eight is CS/70/EXT.

If the base network is defined as the maximum expansion (i.e. fully expanded network containing all proposed projects) of the present CTA rail network, then transit network patronage and consumer surplus changes due to each link investment can be determined by subtracting the link from the maximal system one at a time. This approach also has 8 experimental cases. The abbreviations for this approach are D/30/MAX, D/40/MAX, ..., CS/30/MAX, CS/40/MAX, ..., CS/70/MAX.

These 16 cases are the basis for the first round experiment which is the test of interdependence.

Table 4.2 shows the optimization criteria quantified by the two base network approaches. These values are introduced into the Integer Programming model as the coefficients of the objective function.

Alternative Number	Network Demand Based on*		Consumer Surplus Based on**		Cost ***
	Ext. Sys.	Max. Sys.	Ext. Sys.	Max. Sys.	
1	653	580	41656.	36097.	70
2	2329	1479	179636.	124783.	290
3	2972	1723	259952.	184636.	330
4	3447	1931	194710.	115172.	295
5	1472	459	114044.	51307.	155
6	847	736	78671.	65112.	160
7	902	508	66939.	48483.	150
8	2327	1536	182299.	141628.	230
9	2419	2182	268578.	251433.	350

* Net increase of one-way trips

** Travel time savings in minutes

*** Capital investment in millions of dollars

Table 4.2 Marginal Benefits and Cost for Each Alternative

From Table 4.2, we can observe that optimization criteria quantified by project addition to the existing network system are always higher than those quantified by deletion from the fully expanded (i.e. maximal) network system. It was proven in the previous chapter that if proposed links are competing, the marginal link-specific benefit measures based on the existing system are always higher (upper bounds) than the marginal link-specific measures based on the maximum network system. From Table 4.2, the net total transit patronage increase by adding alternative link #4 to the existing CTA rail network is 3447 passengers and that for alternative Link #5 is 1472 passengers. If we can assume that these two links are independent, the net total transit patronage increase by adding these two to present CTA network system could be 4919 passengers. But the actual patronage increase of the new network having those two alternatives is only 4128

passengers which is obtained from the demand analysis by simultaneously adding those two links. Also note that the demand estimations based on the maximal network are 1931 for alternative Link #4 and 459 for alternative Link #5 and the demand decrease from the joint deletion is only 2390. This is the real world interdependence effect which has to be accounted by the implicit enumeration methods.

In order to examine this interdependence effect on network optimization, the Integer Linear Programming model was used without any modification of the link-specific benefit estimates in Table 4.2. Table 4.3 shows the summary of the optimal solutions for the sixteen cases.

From Table 4.3, it is clear that the Integer Linear Programming model identifies different solutions for most of the cases. This is due to the two different base network assumptions from which the link-specific benefits are determined. As an example, the optimal solution for the D/50/EXT case is Links 3,4,5 and 8 while the optimal solution for D/50/MAX case is Links 1,4,8, and 9. The network composed of Links 3,4,5 and 8 has 1.8 percent less demand than the network composed of Links 1,4,8, and 9. Among 8 comparative cases, the maximum base network approach is better in 3 cases than the existing base network with no difference appearing in another 3 cases. However, this is not sufficient evidence to conclude which is the better approach.

In order to find out how close this integer programming solution is to the global optimal solution(s), a complete search was conducted by enumerating all feasible combinations of the 9 alternative links. Figure 4.1 shows all feasible combinations of the 9 alternative links. Figure 4.1 shows all feasible combinatorial network evaluations under a budget of 812 million dollars, which is 40 percent of 2.03 billion.

Point number 2 is the optimal network plan obtained from the existing base network approach and point number 51 is the optimal network plan obtained from the maximum base network approach. It is now obvious that neither of these two base network approaches (i.e. EXT or MAX) can provide the global optimal plan that is represented by point number 3 in Figure 4.1.

Table 4.3 Comparison of Link-Specific Measurement Approaches

Cases	Optimal Alt. Selected	Budget Left Over	Obj. Func. Value	Actual Benefit from Simultaneous Analysis
D/30/EXT D/30/MAX	1,4,8 1,4,8	\$14 million 14	6427 trips 4047	6204 trips 6204
D/40/EXT D/40/MAX	1,2,4,5 1,6,8,9	2 2	7901 5034	6847 6139
D/50/EXT D/50/MAX	3,4,5,8 1,4,8,9	5 70	10218 6229	8361 8515
D/70/EXT D/70/MAX	1,2,3,4,5,8 1,2,4,6,8,9	51 26	13200 8444	10514 11293
CS/30/EXT CS/30/MAX	8,9 8,9	29 29	450900 ^{min} 393000	448400 ^{min} 448400
CS/40/EXT CS/40/MAX	1,5,8,9 1,6,8,9	7 2	606600 494200	575800 559100
CS/50/EXT CS/50/MAX	1,3,8,9 1,3,8,9	35 35	752500 613700	742300 742300
CS/70/EXT CS/70/MAX	3,4,5,8,9 1,2,3,7,8,9	61 1	1019500 787.00	901500 919500

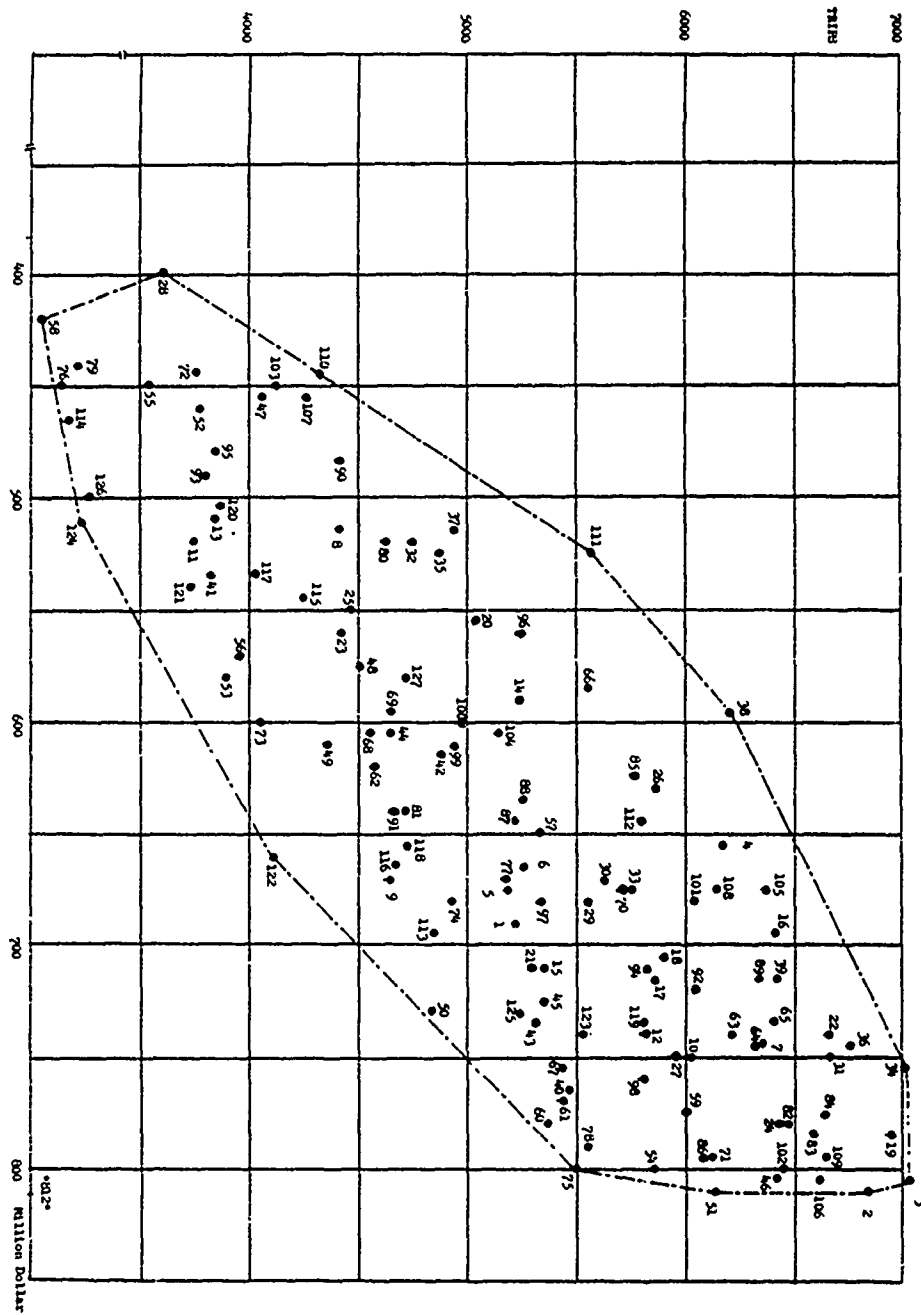


Figure 4.1 Complete Enumeration of Alternative Plans under a Budget of 812 Million Dollar

4.3. Efficiency Test of the Implicit Enumeration Methods

The question raised in this part concerns the efficiency of the branch and bound or backtrack algorithms. Is the necessary number of network evaluations exponentially increasing as the number of proposed links are increased? In order to address this question the network optimization problem for D/40/xxx, i.e. the design problem of maximizing total transit patronage subject to a 812 million dollar budget constraint, is selected because the complete enumeration results are available for this case.

First the Scott backtrack algorithm modified with the linear approximation technique is applied. This backtrack programming algorithm requires a total of 15 network evaluations out of 259 feasible solutions,

When the modified Ochoa-Rosso and Silva algorithm with a linear approximation technique is applied, the total number of network evaluations by the network evaluation program package(RTAEVAL) is only 11. It is interesting to observe that the Ochoa-Rosso and Silva algorithm finds the first feasible solution from the nodes in the level five of the Scott algorithm and that the number of network evaluations required for the Ochoa-Rosso and Silva algorithm is less than that of Scott. The number of network evaluations required for the modified Ochoa-Rosso and Silva's method by the linear approximation technique increases from 3 for the four link problem to 11 for the eleven link problem. However, if the linear approximation technique is not applied in conjunction with the Ochoa-Rosso and Silva's method, the use of their branch and bound approach requires 76 network evaluations. As a note, if the linear approximation method is not applied in conjunction with Scott's backtrack algorithm, use of this approach for the nine link problem(i.e. D/40/xxx) requires 122 network evaluations(i.e. one network evaluation for level three and 121 for level four).

A summary of the efficiency measures for two different size problems is shown in Table 4.4. We are now in a position to compare empirically the relative efficiency of three general methods for attacking combinatorial network design problems. Among the three general approaches, (i.e. the implicit enumeration method, the random sampling method, and the heuristic method) the least practical is a direct implicit enumeration approach(a "brute force" approach in the words of Harris(B. Harris, 1970)) which attempts, through some vari-

ation of the branch and bound method or backtracking, to perform the equivalent of a complete enumeration. This approach totally ignores the possibility of approximating the network benefit (or the non-linear objective function) in a relatively efficient way and so turns out to be an exponential time algorithm. However, it is a very encouraging finding that a modification of this approach by the linear approximation technique can reduce the computational demand (i.e. the number of network evaluations actually required) and still find the globally optimal solution.

Method	N = 4	N = 9	Remarks
Implicit Enumeration Methods	7 - 8	76 - 122	Exponential Increase The globally optimal solution can always be found.
Implicit Enumeration Method Modified by the Linear Approximation Technique	3 - 4	11 - 12	Non-exponential Increase The globally optimal solution can always be found
6.7 Percent Random Sampling Method	1	35	Exponential Increase 50 percent chance of finding one of the best ten solutions
Heuristic Methods Using Recursively Linear Integer Programming Model	4	9 - 27	Polynomial Increase (Less than N^2) Non-dominated solutions, local optimal solutions, or the globally optimal solution are found depending upon the particular case

Table 4.4 Number of Network Evaluation Required for Each Method

5. CONCLUSION

The research described in this paper is a first attempt to develop a public transit network optimization method using a joint destination and mode choice model. The empirically examined modelling framework has permitted an efficient search over the discrete (combinatorial) decision variable space identified for public transit network investment planning. In detail, the results of this study are reviewed under two categories: (a) efficient network benefit quantification based on behavioral travel demand analysis, and (b) efficient network optimization based on the implicit enumeration methods. In the first category the advantages of using a joint choice demand model are explained theoretically and then examined empirically based on a Chicago area application. The review of the second category is based on results obtained from the Chicago area case study.

The theoretical basis for joint choice modelling is that the choice of a place to live and transportation mode to work is not made independently but made jointly based on the relative weights of the living environment conditions and the level of transportation service. One advantage of this simultaneous modelling approach, besides the theoretical and efficiency reasons, is that consistent level of service variables can be introduced linking the trip generation step to the link assignment step. In other words, there is no need to estimate separately the travel times and costs of the trip generation and distribution phase from the mode choice analysis. Depending upon the public transit network changes, a minimum disutility path algorithm can identify the interzonal travel time and cost that can be used in the joint choice modelling analysis.

The development of a public transit network optimization method which can efficiently apply the chosen joint choice model comprises the key part of this research. As stated in Chapter 1, the most critical question is the efficiency of the installation of behavioral demand models in nonlinear discrete network optimization procedures. First, the Branch and Bound, or Backtrack, techniques for network optimization were examined and the idea of using the marginal (or link-specific) benefit measures was presented. It was demonstrated that to improve the efficiency of the implicit enumeration methods the use of link-specific benefit measures was necessary for the branch rejection test. If the linear sum of the link-specific (or marginal) benefit measures is always bigger

than or equal to the actual network benefit (i.e. $L(X) = \sum_1 B_1 \cdot X_1$ $Z(X)$), then the globally optimal solution can be found efficiently. It is proved mathematically that the necessary condition for $\sum_1 B_1 X_1 \geq Z(X)$ is $\text{MIN}(t_{ijk}, t_{ijk}, \dots, t_{ijk}) \leq (t_{ijk}^{A,B,\dots,N})$

, where A, B, \dots, N are the proposed alternatives (lines or links). In other words, if the proposed alternatives A, B, \dots, N compete for passengers, the upper bound marginal (or link-specific) benefits can be obtained for every A, B, \dots, N so that the linear sum of those upper bound marginal benefits is always bigger than the actual network benefit.

For the Chicago Metropolitan Area Transit Network Design problem, the implicit enumeration methods modified based on the linear approximation technique could efficiently search out the globally optimal solution. The nine corridor rapid transit line investment problem clearly demonstrated that the modified implicit enumeration method could be more efficient than a 6.7 percent random sampling approach. The 6.7 percent random sampling method guarantees only a probability of selecting one of the best ten networks.

The important finding from the case study is that the discrete network optimization problem can be solved efficiently (better than a 6.7 percent random sampling approach) without ignoring the trip makers' destination, mode and link choice behavior. The proper procedure of using this promising tool for multi-objective optimization according to alternative futures planning concepts is left for future research.

REFERENCES

- (1) Ben-Akiva, M., Structure of Passenger Travel Demand Models, Ph.D. Dissertation, Department of Civil Engineering, MIT, Cambridge, Mass., 1973.
- (2) Boyce, D., Farhi, A., and Weschiedel, R., "Optimal Network Problem: A Branch-and-Bound Algorithm," Environment and Planning 5, 1973, 519-533.
- (3) Boyce, D. and Seberanes, J., "Solutions to the Optimal Network Problem with Shipments Related to Transportation Cost," Network Design Workshop Paper, Department of Civil Engineering, University of Illinois at Urbana-Champaign, Urbana, Illinois, 1977.
- (4) Brand, D., "The State of the Art of Travel Demand Forecasting: A Critical Review," Paper presented Williamsburg Conference on Travel Demand Forecasting, Williamsburg, Virginia, 1972.

——, "Least-Squares Estimation of Trip Distribution Parameters," Transportation Research 9, February, 1975, 13-18.
- (5) Cesarie F.J., "Trip Generation and Distribution: The Inconsistency Problem and A Possible Remedy," Transportation Planning and Technology 4, 1977, 57-62.
- (6) Harris, B., "Generating Projects for Urban Research," Environment and Planning 2, 1970, 1-21.
- (7) Hutchinsen, B.G., Principles of Urban Transport Systems Planning, Scripts Book, Washington, D.C., 1974.
- (8) Leblanc, L., "An Algorithm for the Discrete Network Design Problem," Transportation Science 9, No.3, 1975, 183-199.
- (9) Lee, D.B., "Requiem for Large Scale Models," Journal of the American Institute of Planners 39, No.3, 1973, 163-178.

- (10) Lerman, S.R., A Disaggregate Behavioral Model of Urban Mobility Decisions, Ph.D. Dissertation, Department of Civil Engineering, MIT, Cambridge, Mass., 1975.
- (11) Lisco, T.E., The Value of Commuter's Travel Time: A Study in Urban Transportation, Ph.D. Dissertation, Department of Economics, University of Chicago, Illinois, 1967.
- (12) McFadden, D., "Conditional Logit Analysis of Qualitative Choice Behavior," in Frontiers in Econometrics edited by Zarembka and Paul, Academic Press, New York, 1973.
- (13) Ochoa-Rosso, F., Application of Discrete Optimization Techniques to Capital Investment and Network Synthesis Problems, Ph.D. Dissertation, Department of Civil Engineering, MIT, Cambridge, Mass., 1968.
- (14) Ridley, T.M., "An Investment Policy to Reduce Travel time in a Transportation Network," Operation Research Center Report 34, University of California, Berkeley, California, 1965.
- (15) Roberts, P.O., "Forecasting Long-Range Travel Demand for Urban Transportation Facilities," Paper presented at the HRB Urban Travel Demand Forecasting Conference, Williamsburg, Virginia, 1972.
- (16) Scott, A., "The Optimal Network Problem: Some Computational Procedures," Transportation Research 3, 1969, 201-210.
- (17) Steenbrink, P., Optimization of Transport Networks, Wiley Beck, London, 1974.
- (18) Stopher, P.R. and Lisco, T.E., "Modelling Travel Demand: A Disaggregate Behavioral Approach: Issues and Applications," Transportation Research Forum Proceedings, 1970.
- (19) Wilson, A.G., Urban and Regional Models in Geography and Planning, John Wiley & Sons, London and New York, 1974.

SOLVING A DISTRIBUTION PROBLEM
WITH DANTZIG-WOLFE DECOMPOSITION:
A CASE STUDY

LUDO F. GELDERS and TONY J. VAN ROY

Katholieke Universiteit Leuven
Celestijnenlaan 300 B
B-3030 Leuven-Heverlee, BELGIUM

ABSTRACT. The purpose of this paper is to present a real-life distribution case study. The problem relates to the distribution of a bottled product from the Antwerp port area over the Belgian territory. The commodity is shipped to depots by company-owned trucks. Then the local distributors take care of the distribution to different consumer areas.

The purpose was to build a distribution system which allows to minimize overall costs, i.e. the sum of transportation, inventory and depot location costs.

The global model allows for the determination of depot location and the allocation of distributors to individual depots. The transportation being based upon trucks, it was impossible to model this situation as a networkflow problem. Moreover specific social constraints (e.g. time schedules of truck drivers) had to be added to the model.

In this paper we concentrate upon the problem of how to allocate consumer areas to fixed depots. A solution procedure based upon Dantzig-Wolfe decomposition was implemented.

1. PROBLEM DESCRIPTION

This paper concerns the distribution of a liquid bottled product from the Antwerp plant over the Belgian territory. The commodity is shipped to 3 depots and 4 main distributors by company-owned trucks (primary transportation). Distribution from these 7 locations to 40 different customer areas is organized by regional distributors or independent contractors. The company pays for the depot and the secondary transportation according to contract. Distribution from the regional distributors to the 2000 individual dealers is organized by the distributors at tariff prices. Figure 1 represents the overall distribution structure.

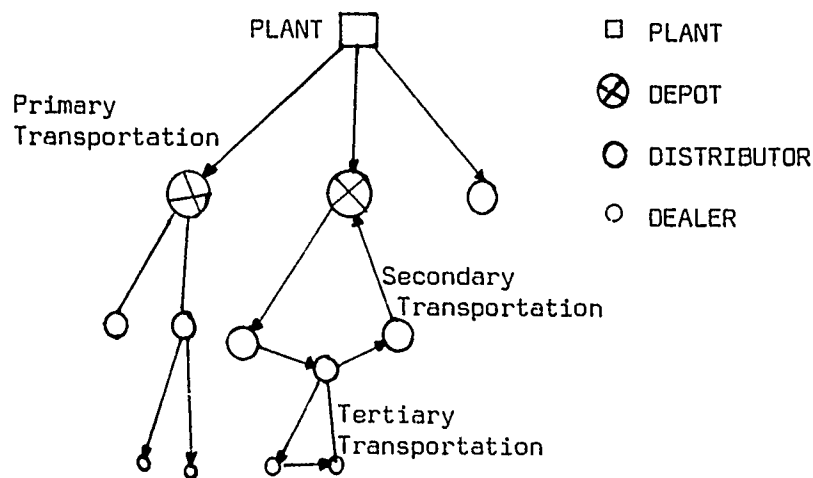


Fig. 1. Distribution Structure

The above distribution structure involves transportation costs, depot operating costs and inventory costs. Total demand and demand pattern being variable in time, the company felt the need to develop a normative model for locating its depots and allocating the distributors to depots on a cost-minimization basis. A variety of constraints (e.g. maximum capacities, truck driver schedules, minimum stocklevels) must be satisfied.

This paper deals with the customers allocation problem, i.e. the allocation of customers to a given set of depots (fixed locations). It results in the implementation of generalized linear programming (Dantzig-Wolfe decomposition)

where the subproblem is a minimum cost network flow problem.

The company operated with 7 heavy trucks and 11 containerized trailers. Taking into account trailer loading operations and truck overhaul, we may consider that 6 loaded lorries are always available for primary transportation. The trucks being slightly different, we introduced a hybrid truck with average capacity. Truck drivers are paid on an hourly basis, overtime being limited by law. Delivery requirements (capacity and distance) could be met with 3 trucks on double full shift (2x8 hours) and 3 trucks on extended single shift (1x11 hours or 2x5.5 hours). Total crew size was 10. The primary transportation cost was divided in a fixed cost and a variable cost (per km and per ton transported). Regular truck driver time was included in the fixed cost.

The secondary transportation fee is based upon a fixed standard (as a function of distance between the depot and the distributory points). As a consequence the distributors or independent transportation contractors may take considerable advantage of choosing the most appropriate truck and/or tour (see Fig. 2).

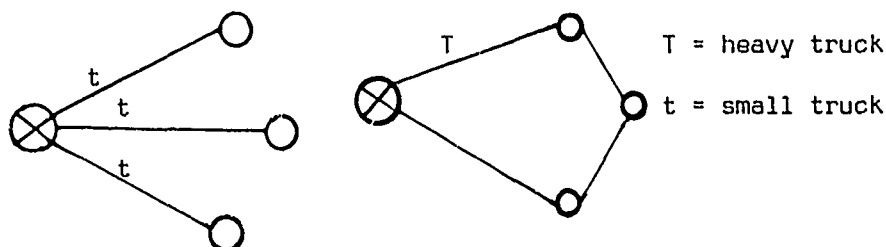


Fig. 2. Secondary Transportation

The distributor receives a fixed fee per ton distributed to the individual dealers. Under this system, the cost of tertiary transportation is not affected by our policy variables, i.e. depot location and secondary transportation allocation.

Three payment schemes for depot owners are used. Under the first scheme they get a fixed annual depot premium plus a small fee per ton throughput. Under the second scheme, the premium depends only on throughput, while in the third scheme the company guarantees a minimum throughput (see Fig.

3). Some of the depot throughputs are constrained.

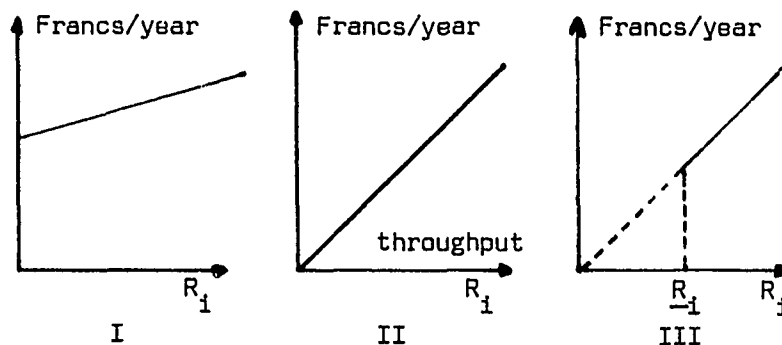


Fig. 3. Depot Holders Premium

The product is distributed in company-owned standardized bottles, which are returned to the company. Loading a truck implies replacing empty bottles by filled bottles or vice-versa. Hence, the inventory level is equal to the truck capacity and the reorder point quantity, which equals $2 \times$ daily throughput for operational reasons. Therefore, the inventory cost may be considered to be proportional to the depot throughput.

The analysis of the cost structure yielded several dominance relations with respect to transportation strategies. Fig. 4 shows that strategy A is better than strategy B if the depot location is rather far away from the plant, the depot operating costs and inventory cost in A being outweighed by the additional transportation cost in B. Strategies C and D, however, are not relevant. They are clearly dominated by strategies C' and D' respectively, provided that the heavy truck T may reach all destinations in a 1 day round-trip.

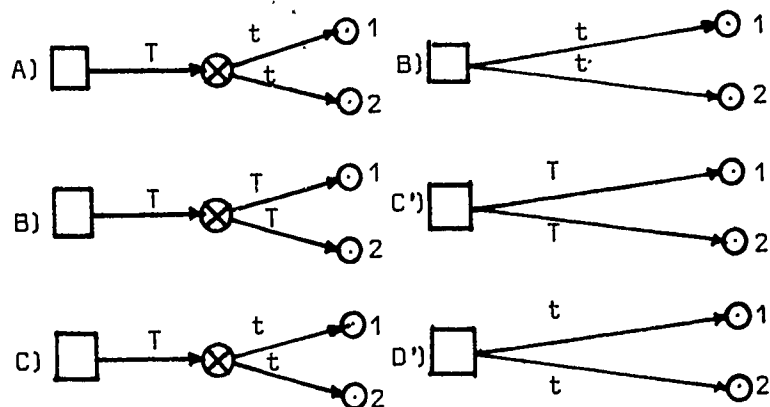


Fig. 4. Transportation Strategies

Finally, it should be mentioned that--given the shift structure mentioned above--the Belgian territory was subdivided into 3 zones (see Fig. 5):

- zone A which may be reached in 5.5 hours round-trip (loading and unloading included). Locations belonging to this zone may be reached two times a day in an extended 11-hours shift;
- zone B including points which may be reached in a round-trip time between 5.5 hours and 8 hours;
- zone C including points which may only be reached once a day with an extended 11 hours shift.

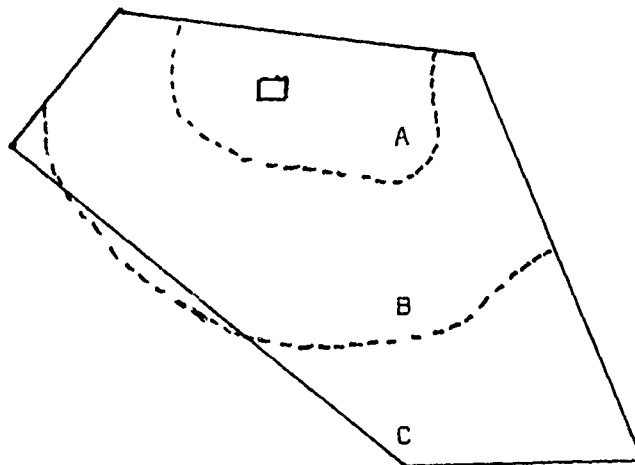


Fig. 5. Zoning

2. THE DECISION MODEL

Following model was used to represent the general decision problem described above (i.e. the combined location-allocation problem):

$$\begin{aligned} \text{Min } \{ & F + \sum_{i=1}^m f_i x_i + \sum_{i=1}^m c_i R_i + (\sum_{i=1}^m c_{1i} R_{1i} + c_u U) + \sum_{i=1}^m c_{2i} R_{2i} \\ & + \sum_{i=1}^m \sum_{j=1}^m c_{ij} r_{ij} \} \end{aligned} \quad (1)$$

subject to:

$$R_i = \sum_j r_{ij} \quad \text{all } i \quad (2)$$

$$x_i R_i \leq R_i \leq x_i \bar{R}_i \quad \text{all } i \quad (3)$$

$$r_j = \sum_i r_{ij} \quad \text{all } j \quad (4)$$

$$R_i = R_{1i} + R_{2i} \quad \text{all } i \quad (5)$$

$$\sum_i k_{1i} R_{1i} \leq K_1 \quad (6a)$$

$$\sum_i k_{2i} R_{2i} \leq K_2 \quad (6b)$$

$$\sum_i t_i R_{1i} - u \leq U \quad (6c)$$

$$u \leq u_{\max} \quad (6d)$$

$$x_i \in \{0, 1\} \quad r_{ij}, R_i, R_{1i}, R_{2i}, u \geq 0 \quad (7)$$

Following symbols are used:

- F annual fixed cost of truck fleet and driver wages
 i depot index (i=1...m)
 j customer area or distributor index (j=1...m)
 m number of candidate depots
 n number of customer areas
 f_i annual fixed depot and inventory cost
 x_i = 1 } if depot is open
 = 0 } if depot is closed
 R_i annual throughput depot i
 c_i variable cost depot i
 R_{1i}, R_{2i} annual volume transported from plant to depot i
 by trucks under the extended single shift system
 and the double full shift system
 c_{1i}, c_{2i} variable primary transportation cost containing
 variable truck and overtime cost
 r_{ij} annual consumption quantity in area j delivered
 through depot i
 c_{ij} secondary transportation cost (from depot i to
 area j)
 u total amount of extra hours per year in the ex-
 tended single shift system, i.e. total amount of
 surplus hours above U
 c_u regular salary (per hour)
 R_i, \bar{R}_i minimum, maximum throughput in depot i
 r_j annual requirement of customer area j
 k_{1i} = 1/2 } for depot i belonging to zone A
 = 1 } for depot i belonging to zone B or C
 K₁ total number of truck trips per year in the ex-
 tended single shift system = number of workdays
 per year x number of trucks

- k_{2i} = 1 for depot i belonging to zones A and B
 k_{2i} = ∞ for depot i belonging to zone C
 K_2 total number of truck trips per year in the double shift system = 2 x number of workdays x number of trucks
 t_i round trip time for primary transportation to depot i (including break time when visiting a depot in zone A)
 u_{\max} maximum amount of hours per year in the extended single shift system
 U total amount of regular hours (e.g. on a 40 hours per week basis) per year.

The terms of the objective function (1) represent consecutively the fixed primary transportation cost, the fixed depot cost, the variable depot cost, the variable primary transportation cost for the extended single shift system, the variable primary transportation costs for the double full shift system, the secondary transportation costs.

Notice that directly provisioned areas are treated as depots which can serve only one customer area. Notice also that extra hours effectively performed during overtime (i.e. after regular hours on a particular day), are automatically accounted for in c_{1i} as far as the additional overtime salary is concerned; the regular salary is accounted for in c_u .

Constraints (2), (4) and (5) represent the flow conservation equations for respectively the depot throughputs, the customer requirements and the primary transportation. The depot throughput has a lower and an upper bound as expressed in (3). Constraints (6a) and (6b) express the truck fleet capacity limitations of primary transportation due to scheduling and zoning considerations. Finally, the upper bound on the total amount of hours per year traveled by drivers of the extended single shift system is represented by constraints (6c) and (6d).

For simplicity, we write constraints (6a)-(6d) into a more general form:

$$\sum_i d_{1i} R_{1i} + \sum_i d_{2i} R_{2i} + d_u U \leq D \quad (6)$$

d_{1i} , d_{2i} , d_u and D being column vectors of appropriate length.

Finding the optimal values of flows r_{ij} of problem (1)→

(7) is called the customer allocation problem or the transshipment problem, while determining the x_i values is known as the depot (facility, warehouse) location problem.

Although this decision model is a (general) mixed integer linear programming model, the straightforward implementation of general codes such as MPSX/MIP, APEX, OPHELIE, would be too expensive (memory requirements and computational times) in comparison with special codes. For example, Glover [3], [4] reports that state-of-art network codes are more than 200 times faster than general linear programming codes for solving real-life customer allocation problems (i.e. transshipment problems).

Although the cost of data collection, problem analysis and problem formulation are independent of the specific algorithm used, it is clear that computational cost remains very important. For example, besides solving the "base case" of our Bottles Distribution Problem, we had to solve about 30 other cases for purposes of sensitivity analysis, trade-off analysis, priority analysis, etc. (See [2] for a detailed discussion of experimental design in logistics distribution system planning).

3. SOLVING THE CUSTOMER ALLOCATION PROBLEM

The existence of constraints (6) prevents a straightforward implementation of a standard transshipment package. Maier [7] described a solution procedure for the "transshipment problem with additional constraints" using a compact inverse scheme and tree labeling techniques, but computational experience was not reported. Recently Glover and Klingman [5] presented the SON package which uses these techniques on a general "di-graph" formulation of the problem.

We turned to the Dantzig-Wolfe decomposition principle for linear systems [6],[8]. Note that the small number of additional constraints could indicate an efficient implementation of this principle.

The linear model represented above may be decomposed into a subprogram and a master program. The price paid for this decomposition is that the master and the subprogram may have to be solved several times. First the master program is solved, and from its solution, the objective function is generated for the subprogram. Then this problem is solved, and from its solution a new column is generated to be added to the master program.

More formally the Dantzig-Wolfe decomposition of problem

(1) + (7), x_1 fixed, yields the following master problem, the fixed terms of (1) being dropped:

$$\text{Min } \sum_{t=1}^T \lambda_t p^t + c_u u \quad (1')$$

$$\text{s.t. } \sum_t \lambda_t \{ \sum_i d_{1i} R_{1i}^t + \sum_i d_{2i} R_{2i}^t \} + d_u u \leq D \quad (6')$$

$$\sum_t \lambda_t = 1 ; \lambda_t \geq 0 \quad \text{all } t \quad (8)$$

with one subproblem:

$$\begin{aligned} \text{Min } & \sum_i c_i R_i + \sum_i (c_{1i} - \pi^t d_{1i}) R_{1i} + \sum_i (c_{2i} - \pi^t d_{2i}) R_{2i} \\ & + \sum_i \sum_j c_{ij} r_{ij} \end{aligned} \quad (1'')$$

$$\text{s.t. } (2), (3), (4), (5), (7)$$

$$x_1 \text{ fixed}$$

where t is the index of the iteration;

p^t is the cost of the solution at iteration t ;

π^t denotes the row vector of shadow prices of the constraints (6);

(8) is the convexity constraint with λ_t the variable of the t^{th} iteration.

The master problem is a general LP problem, solved by the simplex algorithm. The subproblem is basically the transshipment problem represented by the network of Fig. 6.

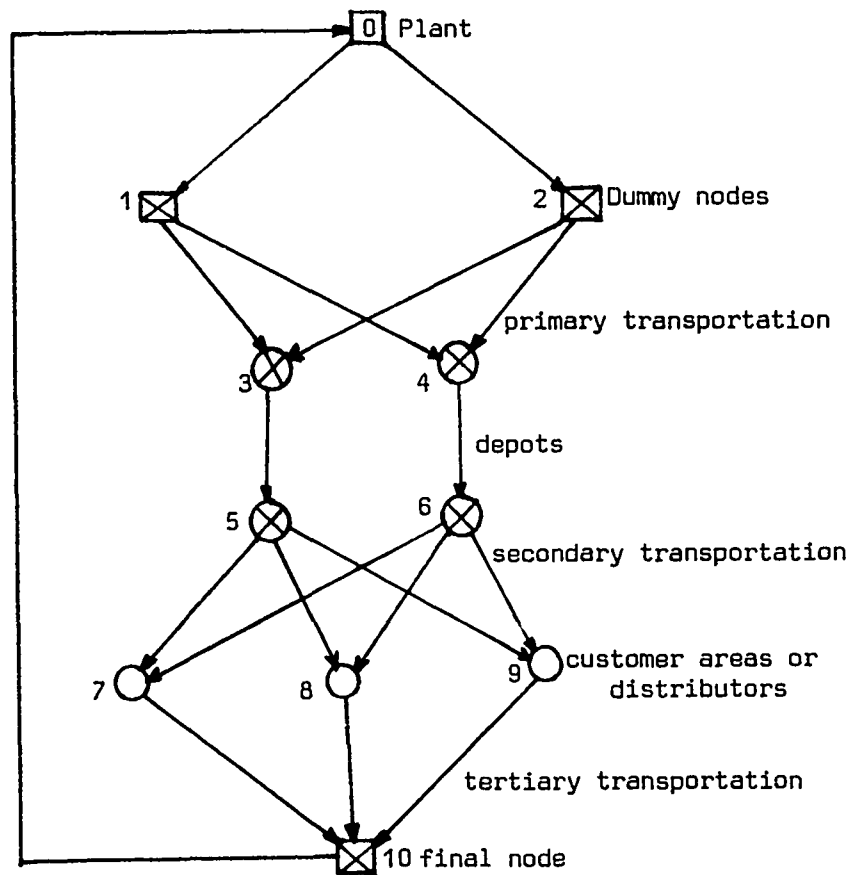


Fig. 6. Network Presentation of the Problem without the Additional Constraints

In this network nodes represent the plant, the depots and the distributors respectively. The flows of the primary transportation arcs originating from dummy nodes 1 and 2 are represented by R_{11} and R_{21} respectively. Depot throughput constraints of type (3) are applicable on the depots arcs (3, 5) and (4, 6).

Maximum distributor demand yields an upper bound on secondary transportation arcs, the lower bound being equal to zero. The same rule applies to the tertiary transportation arcs. Note that some, but not all, of the constraints of type (6) can be presented by the network. For example constraint (6b)

can be written as:

$$\sum_{i \in AB} R_{2i} \leq K_2, \quad R_{2i} = 0 \text{ for } i \notin AB$$

where AB is the set of all depots with $k_{2i} = 1$, i.e. belonging to zones A and B. In the transshipment network we introduce an upper bound on arc (0, 1), and drop all the arcs originating in node 2 and going to depot nodes of zone C.

4. COMPUTATIONAL RESULTS

The solution procedure described above was coded in FORTRAN IV. The master problem of the distribution problem containing only 4 constraints, we used an explicit inverse revised simplex code. For contractual reasons an out-of-kilter code was used for solving the transshipment problem, although more efficient codes were available (see e.g. [1]). The actual version may solve problems with 10 additional constraints of type (6), 1000 nodes and 2000 arcs. The Bottles Distribution Problem has 7 depots, 40 customers and 3 general linear constraints; computation time varied from 3.5 to 4.5 CPU seconds on an IBM 370/158 for solving one problem, including input and output; a maximum of 5 iterations had to be performed.

5. REFERENCES

- [1] Bradley, G., Brown, G.G. and Graves, G.W., DESIGN AND IMPLEMENTATION OF A LARGE SCALE PRIMAL TRANSSHIPMENT ALGORITHM, Management Science, Vol. 24, NO 1, PP. 1-34, 1977.
- [2] Geoffrion, A.M. and Graves, G.W., MULTICOMMODITY DISTRIBUTION SYSTEM DESIGN BY BENDERS DECOMPOSITION, Management Science Vol. 20, NO 5, PP. 824-844, 1974.
- [3] Glover, F., Karney, D., Klingman, D. and Napier, A., A COMPUTATION STUDY ON START PROCEDURES, BASIC CHANGE CRITERIA, AND SOLUTION ALGORITHMS FOR TRANSPORTATION PROBLEMS, Management Science Vol. 20, NO 5, PP. 793-813, 1974.
- [4] Glover, F. and Klingman, D., NETWORK APPLICATIONS IN INDUSTRY AND GOVERNMENT, AIIE Transactions, Vol. 9, NO 4,

PP. 363-376, 1977.

- [5] Glover, F. and Klingman, D., THE SIMPLEX SON ALGORITHM FOR LP/EMBEDDED NETWORK PROBLEMS, Research Report CCS 317, Center for Cybernetic Studies, The University of Texas at Austin, 1977.
- [6] Lasdon, L.S., OPTIMIZATION THEORY FOR LARGE SYSTEMS, Macmillan, London, 1970.
- [7] Maier, S.F., A COMPACT SCHEME APPLIED TO A MULTICOMMODITY NETWORK WITH RESOURCE CONSTRAINTS, in: Cottle, R.W. and Krarup, J. (eds.), Optimization Method for Resource Allocation, The English University Press, London, 1974.
- [8] Weigel, H.S. and Cremeans, J.E., THE MULTICOMMODITY NETWORK FLOW MODEL REVISED TO INCLUDE VEHICLE PER TIME PERIOD AND NODE CONSTRAINTS, Naval Research Logistics Quarterly, Vol. 19, NO 1, PP. 77-89, 1972.

THE DEVELOPMENT OF FERTILIZER DISTRIBUTION SYSTEM
-AN APPLICATION OF THE TRANSPORTATION
LINEAR PROGRAMMING MODEL-

YOON, YONG WOON

Department of Industrial Mgt.
Mokpo Technical College
#525, Sangdong, Mokpo City
Chunnam-Province, Korea

THE PACIFIC CONFERENCE ON OPERATIONS RESEARCH
OR in Public Administration
(Transportation Systems Analysis)

Dec. 28, 1978

THE DEVELOPMENT OF FERTILIZER DISTRIBUTION SYSTEM
-AN APPLICATION OF THE TRANSPORTATION
LINEAR PROGRAMMING MODEL-

YOON, YONG WOON

Department of Industrial Mgt.
Mokpo Technical College
#525, Sangdong, Mokpo City
Chunnam-Province, Korea

ABSTRACT. The purpose of this study is to establish the optimum model of transportation system by transportation methods for fertilizer, and develop a fertilizer distribution system with regard to the shortening transportation distances and the savings of transportation cost which are resulted from a new fertilizer distribution model in order to control harmoniously the demand-supply and keep up the optimal inventory level for fertilizer with the conversion of the free sale policy for fertilizer. That is to say, I analyze transportation linear programming technique which is the transportation model minimizing the transportation cost between supply area and consumption area on the restricted conditions of transportation capacity, freight working and custody scale for consumption area and transportation methods and I was established the best proper transportation model to this with the supply and demand record of fertilizer in 1975, compare and evaluate it through simulation with the record and studied the basic direction of general management system including the transportation management and supply and demand management by the economic transportation model establishment.

This study is the matter which was performed as assistant project of Ministry of Science & Technology. I announce that there was Mr. Yoon's assistance who is Director of Economic Research Department, KID for this study.

1. INTRODUCTION

1.1. Purpose of the Study

The purpose of this study is to suggest a model of the fertilizer distribution system which makes managers better use of the inventory and distribution network system on the supply and demand for fertilizer.

1.2. Scope of the Study

The study has been analyzed mainly based on the current existing distribution system in order to develop a total supply-demand system.

- (1). Analysis of the supply-demand for fertilizer in 139 City and Kun is done (Except for Ongjin-Kun, Gyeounggi-Province, Ulreung-Kun, Gyeoungsangbug-Province) as shown in Table 1.

We analyze transportation linear programming technique which is the transportation model minimizing the transportation cost between supply area and consumption area on the restricted conditions of transportation capacity, freight working and custody scale for consumption area and transportation methods and we established the best proper transportation model to this with the supply and demand record of fertilizer in 1975, compare and evaluate it through simulation with the record and studied the basic direction of general management system including the transportation management and supply and demand management by the economic transportation model establishment.

- (2). An evaluation study is done to measure the effects on the shortening transportation distances and the savings of transportation cost which are resulted from a new fertilizer distribution model.

Especially, as the selection of the first fertilizer transportation destination in the progress method of reasonable transportation system by the transportation model establishment is the important variable figure of transportation cost accounting by the transportation cost pool system,¹ we took the scope

1. Pool system of transportation cost a system of accounting

with the weighted freight transportation distance accounting to the final destination and introduced it to the best proper transportation distance model establishment through the transportation relay place from the respective City and Kun unit warehouse zone in this study.

- (3). Recommendations are made based on the results of the study to make a better fertilizer distribution system.

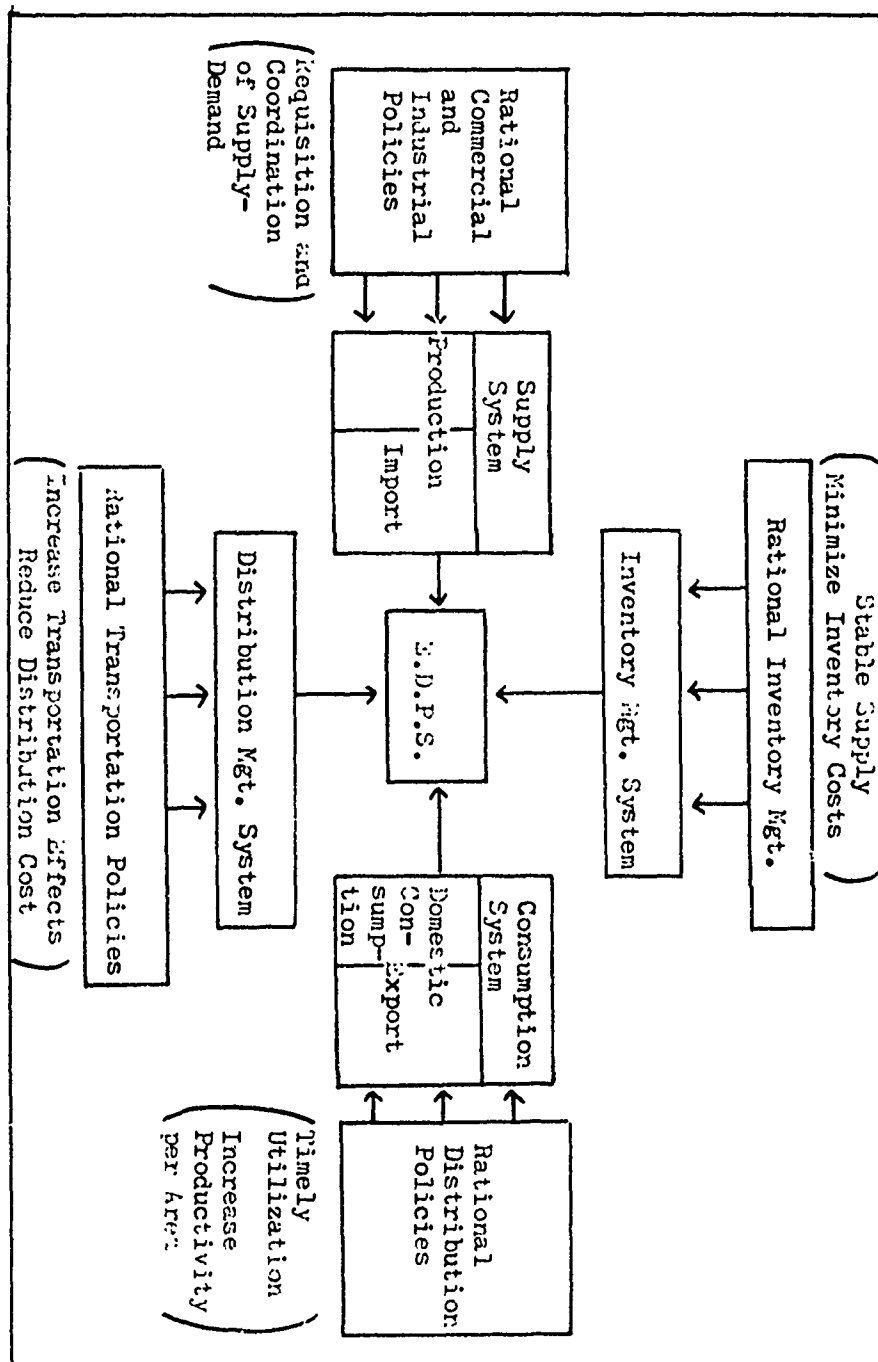
1.3. Expansion of the Study

In the first case, we took the circulation activities only which are the connective function between supply and consumption area as the following matters should be enforced for its realization and the comparative evaluation should be accomplished through the continual simulation of the fertilizer supply and demand system model for its achievement.

- (1). As we develop the estimated demand model of fertilizer to the respective unit warehouse and the consumption area, general supply and demand management system model and the connective function should be achieved.
- (2). Proper selection of transportation relay place and making of relay place by this selection and decision of the required investment scale.
- (3). As we develop the productive control management system which is able to control the freight capacity by the demand change to such transportation relay place and consumption area (final destination).
- (4). And we should fix the proper selective system of the alternative method which can be applied to the fertilizer supply and demand circumstances change.

and paying the total transportation cost with double of fertilizer transportation unit price and total transportation weight which are accounted by the former year transportation record without transportation distance and methods in payment of transportation cost.

1.4. Schematic Diagram for the Total Transportation System



2. THE ANALYSIS OF SUPPLY AND DEMAND FOR FERTILIZER

2.1. Supply and Demand for Fertilizer

2.1.1. Yearly Analysis of Supply and Demand for Fertilizer

2.1.1.1. Production

See Table 2.

2.1.1.2. Consumption by Kinds of Fertilizer

See Table 3.

2.1.1.3. Yearly Consumption

See Table 4.

2.1.2. Geographical Analysis of Supply and Demand for Fertilizer

See Table 5 and Fig. 1.

2.2. Analysis of Distribution Channel for Fertilizer

2.2.1. Analysis of Transportation Methods

See Table 6, Table 7.

2.2.2. Schematic Diagram for Supply of Fertilizer

See Fig. 2.

2.3. Cost Analysis of Transportation for Fertilizer

2.3.1. Transportation Cost Analysis Per Ton by Methods

The scope of transportation cost of fertilizer as the analysis subject in this study is limited to the cost of occurring to the respective transportation methods directly as shown in Table 8.

But as it is impossible for the problem of transportation unit price to the respective place between original and destination to account the individual record value of transportation to the respective place between original and destination owing to the pool system of fertilizer transportation cost, we selected the following accounting method.

2.3.1.1. Railway

Railway transportation unit price to equivalent section means the total transportation cost per each section with railway freight charge $\{ \text{Freight Charge} = \text{Class}(\text{basic charge rate}) \times \text{Transportation Distance}(\text{section number}) \times \text{Weight}(\text{weight accounting freight}) \}$ and incidental cost for transportation altogether. And we counted the section numbers the respective place between original and destination and applied total transportation unit price to the equivalent section.

2.3.1.2. Public Road

Public road transportation unit price by the vehicles puts public road freight charge $(\text{transportation distance} \times \text{basic charge rate})$ and incidental charges for transportation together.

2.3.1.3. Coast Marine Transportation

Coast marine transportation freight charge accounts transportation distance(Nautical mile) between original and destination and we apply and account it to the freight charge basis rate. This basic freight charge rate is the long distance tapering rates and the incidental charge is applied by the basis of freight working charge rate of Ministry of Transportation.

2.3.2. Transportation Cost Analysis Per M/T-Km by Methods

As the calculated results of transportation unit price to the transportation distance of the respective transportation methods as the above, we can know the following features.

First, in case of fertilizer transportation to the transportation area which transportation distance is over 50 Km. we can know that transportation by railroad is favorable, but public road transportation by vehicles is unfavorable relatively. That is to say, transportation charge rate of railroad is same per 50 Km for transportation charge although it is 1 Km or 49 Km by vehicles because the freight charge rate is applied by the section on the basis of 50 Km per a section. But public road transportation charge rate by vehicles is different with transportation charge rate per Km.

In case of public road transportation, transportation charge rate within 6 Km is accounted with the direct distribution charge conception by disposal condition of small transportation freight charge and the accounting of transportation charge rate for over 6 Km of transportation dis-

tance is calculated to double the transportation distance to the transportation charge rate per Km(See Fig. 3).

As transportation unit price per unit(per M/T) to long distance transportation for over 50 Km, we can know that transportation charge rate by railroad is more favorable than that of charge rate by the other transportation methods.

Second, in the view of gravity for incidental cost of transportation in the composition of transportation unit price for the respective transportation methods, coast marine transportation part is high relatively. As the above Table 9, in the view of similar level in transportation unit price without relation of merits and demerits of transportation distance, railroad part is composed with basic freight charge 79.7% of the transportation unit price, but the gravity of incidental cost is calculated to 20.3%. And public road part is calculated 81.2% and 18.8% each. But the transportation unit price per M/T of coast marine transportation part by seashore ships is 49.9% for composition ratio of the basic freight charge and it takes almost half. So it is much difference to other transportation methods level, but the gravity of incidental cost is 50.1% and more two times over than 18% or 20% of transportation incidental cost gravity of railroad and public road. We can know that transportation incidental cost gravity of coast marine transportation is high.

Third, transportation unit price for the respective transportation methods is applied to 50 Km per a section in the railroad case and coast marine transportation is applied to 20 nautical mile or 100 nautical mile as the basic distance per unit section and this applying of transportation charge rate can show features of long distance of transportation and heavy weight of transportation, but the transportation of the special geographic area is impossible in reality. So the distribution rate of transportation by public road must be increased inevitably. Railway transportation is more favorable at large in case of transportation of fertilizer except mountain area or short distance transportation, near sea coast area.

2.3.3. Incidental Cost Analysis for the Respective Transportation Methods

On the other hands, if we see composition of incidental cost for the respective transportation methods, decision of transportation unit price is large on the gravity which is made by the basic transportation charge rate in railway or public road as the above explaining and transportation cost of coast marine transportation is large on the gravity which is made by incidental cost charge rate.

2.3.3.1. Railway

If we see the composition of transportation unit price per M/T in the railroad case, the gravity of incidental cost to the basic freight charge is 20.3% and loading charge is 29.4% in the factory and the primary transportation relay place and unloading charge is 26.8% in the distribution and relay place. So the loading and unloading charge take 56.2% of total incidental cost and damaged sack packing charge is 34.7%.

The above ratio is composed of damaged packing charge to the paper sack and damaged sack packing charge to the damaged part in time of unloading by the laborer and this gravity is 30.8% and 3.9% each of total incidental cost. So it is shown that they are larger than that of loading and unloading charge.

2.3.3.2. Public Road

Composition of incidental cost in the public road transportation by the vehicles is applied by incidental cost in time of railroad transportation, but dispatch cost of factory entry line is excluded. So the gravity of incidental cost in the composition of public road transportation unit price is 18.8% and the gravity of incidental cost of railroad is low, but gravity of the basic freight charge is 81.2% level.

2.3.3.3. Coast Marine Transportation

Gravity of incidental cost in the coast marine transportation takes 50.1% high than that of incidental cost of transportation unit price by any other transportation methods and the basic freight charge is lower than that of any other transportation methods because the coast marine transportation is tapering rate of long distance but the gravity of incidental cost is high. (See Table 9)

3. DEVELOPMENT OF AN OPTIMUM TRANSPORTATION SYSTEM AND AN ECONOMICAL TRANSPORTATION MODEL

3.1. Decision on Transportation Points

3.1.1. Selection of Relay Points

Transportation cost reduction effectiveness by model of transportation path become a main composition factor of the primary transportation relay points selection. Therefore, fertilizer transportation path to the final destination through the optimal transportation relay place from each supply area can be supposed as the following path in this study approximately as shown in Fig.4, Fig. 5.

This constitution of transportation path of the respective transportation methods is to be a judgement basis to keep the proper transportation system and the understanding of proper transportation path of fertilizer is to be able to the best transportation network formation, maximum of transportation cost and reduction effectiveness.

We select the transportation relay place by the transportation share rate of the respective methods as the premise condition for the optimal transportation system and the economical transportation model development of fertilizer, but it was impossible to get the materials to this.

So I selected the transportation relay place by the following selective basis.

- (1). Selection of the primary transportation destination to the nation wide 139 City and Kun calculated the weighted freight transportation distance apart the final destination through the relay place from the unit warehouse zone of the respective City and Kun and selected it as the supposed destination where is located to the shortest transportation distance.
- (2). Selection of relay place which is premise condition of the final transportation destination selected the relay point where required minimum level the transportation cost of the transportation unit price between original-destination to the respective transportation method.
- (3). Selection to the transportation relay place through the complicated transportation path choosed the shortest relay place with supply area and this was possible as comparing the transportation unit price to the respective transportation method.

3.1.2. Selected Relay Points and Transportation Cost

3.1.2.1. Selected Relay Points

Then, I selected the primary transportation relay area by the following selected basis as above. (as shown in Table 10)

3.1.2.2. Transportation Cost Analysis between the Selected Relay Points

The transportation cost of the final destination through the complicated transportation path choosed the cost of transportation methods which required the minimum transportation cost between the respective transportation path as shown in Table 11.

3.2. Development of an Optimum Transportation Network and Economical Network Model

3.2.1. Constraints of the Model

Economic transportation model for circulation management system development of fertilizer established the model which selects the transportation root unconditionally by the high and low transportation roots of 1663 subjecting 12 supply areas and 139 demand areas on the basis of supply and demand record of fertilizer in 1975.

3.2.2. Formular for the Model

Induce of objective function which minimize the constraints of Linear Programming formula and transportation cost in the selection of optimal transportation root by economic transportation model of fertilizer is as follows.

3.2.2.1. Constraints

- (1). Supply quantity of fertilizer in the respective supply area can exceed the supply capacity.

$$\sum x_{ij} \leq S_i \quad \dots (1)$$

(i=1,2,...,12, j=1,2,...,139)

$$\left(\begin{array}{l} x_{ij}; \text{ Supply Capacity in } i \text{ th Supply Areas} \\ S_i; \text{ Supply to } j \text{ th demand Areas from } i \text{ th Supply} \end{array} \right)$$

(Areas or Demand from i th Supply Areas)
 in j th Demand Areas

- (2). Quantity of fertilizer to be purchased from the respective consumption area should not be less than demand quantity of the consumption area.

$$\sum x_{ij} = D_{.j} \quad \dots(2)$$

(i=1,2,...,12, j=1,2,...,139)

(D_{.j} ; Demand Capacity in j th Demand Areas)

3.2.2.2. Objective Function

$$\sum C_{ij} \cdot x_{ij} = Z \text{ (minimize) } \dots(3)$$

(C_{ij} ; Average Transportation Cost to j th Demand Areas from i th Supply Areas)

3.2.3. Establishment of An Optimal Transportation Model

Then, the established components of the optimal transportation model constituted the supply function 12, demand function 139 and the variables 1668 for objective function, the variables premise the constraints for transportation model establishment is as shown in Table 12,13.

3.2.3.1. Supply Capacity by Supply Areas for Fertilizer

See Table 12.

3.2.3.2. Consumption by Elements, Demand Areas for Fertilizer

See Table 13.

3.2.3.3. Model of Demand-Supply Function by Elements of Fertilizer

To arrange the constraints and the objective function, the L.P. Model for determination of optimum transportation quantity is shown in the following.

(1). Nitrogen(N)

$$\begin{aligned} \text{Min(transportation cost)} = & C_{1.1} x_{1.1}^N + C_{1.2} x_{1.2}^N + \dots \\ & + C_{12.139} x_{12.139}^N \quad \dots(4) \end{aligned}$$

$$\begin{array}{l} \text{subject to;} \\ \left(\begin{array}{c} x_{1.1}^N + x_{1.2}^N + \dots + x_{1.139}^N \leq S_1^N \\ x_{2.1}^N + x_{2.2}^N + \dots + x_{2.139}^N \leq S_2^N \\ \vdots \\ x_{12.1}^N + x_{12.2}^N + \dots + x_{12.139}^N \leq S_{12}^N \end{array} \right) \dots (5) \end{array} \quad \begin{array}{l} \text{Supply} \\ \text{Constraints} \end{array}$$

$$\left(\begin{array}{c} x_{1.1}^N + x_{2.1}^N + \dots + x_{12.1}^N = D_1^N \\ x_{1.2}^N + x_{2.2}^N + \dots + x_{12.2}^N = D_2^N \\ \vdots \\ x_{1.139}^N + x_{2.139}^N + \dots + x_{12.139}^N = D_{139}^N \end{array} \right) \dots (6) \quad \begin{array}{l} \text{Demand} \\ \text{Constraints} \end{array}$$

$$x_{ij}^N > 0 \quad (i=1,2,\dots,12, j=1,2,\dots,139) \dots (7)$$

(2). Phosphorus (P_2O_5)

$$\text{Min(transportation cost)} = C_{1.1} x_{1.1}^{P_2O_5} + C_{1.2} x_{1.2}^{P_2O_5} + \dots + C_{12.139} x_{12.139}^{P_2O_5} \dots (4)$$

$$\begin{array}{l} \text{subject to;} \\ \left(\begin{array}{c} x_{1.1}^{P_2O_5} + x_{1.2}^{P_2O_5} + \dots + x_{1.139}^{P_2O_5} \leq S_1^{P_2O_5} \\ x_{2.1}^{P_2O_5} + x_{2.2}^{P_2O_5} + \dots + x_{2.139}^{P_2O_5} \leq S_2^{P_2O_5} \\ \vdots \\ x_{12.1}^{P_2O_5} + x_{12.2}^{P_2O_5} + \dots + x_{12.139}^{P_2O_5} \leq S_{12}^{P_2O_5} \end{array} \right) \dots (5) \end{array}$$

$$\left(\begin{array}{c} x_{1.1}^{P_2O_5} + x_{2.1}^{P_2O_5} + \dots + x_{12.1}^{P_2O_5} = D_1^{P_2O_5} \\ x_{1.2}^{P_2O_5} + x_{2.2}^{P_2O_5} + \dots + x_{12.2}^{P_2O_5} = D_2^{P_2O_5} \\ \vdots \\ \vdots \end{array} \right) \dots (6)$$

$$\left(\begin{array}{c} P_{205} \\ x_{1.139} \end{array} + \begin{array}{c} P_{205} \\ x_{2.139} \end{array} + \dots + \begin{array}{c} P_{205} \\ x_{12.139} \end{array} = \begin{array}{c} P_{205} \\ D_{139} \end{array} \right)$$

$$x_{ij}^{P_{205}} > 0 \quad (i=1,2,\dots,12, j=1,2,\dots,139) \dots (7)'$$

(3). Potash(K_2O)

$$\text{Min(transportation cost)} = C_{1.1}^{K_2O} x_{1.1}^{K_2O} + C_{1.2}^{K_2O} x_{1.2}^{K_2O} + \dots + C_{12.139}^{K_2O} x_{12.139}^{K_2O} \dots (4)''$$

subject to:

$$\left(\begin{array}{cccc} K_2O & K_2O & & K_2O \\ x_{1.1} & + x_{1.2} & + \dots + & x_{1.139} \\ K_2O & K_2O & & K_2O \\ x_{2.1} & + x_{2.2} & + \dots + & x_{2.139} \\ \vdots & \vdots & \ddots & \vdots \\ K_2O & K_2O & & K_2O \\ x_{12.1} & + x_{12.2} & + \dots + & x_{12.139} \end{array} \leq \begin{array}{c} S_1^{K_2O} \\ S_2^{K_2O} \\ \vdots \\ S_{12}^{K_2O} \end{array} \right) \dots (5)''$$

$$\left(\begin{array}{cccc} K_2O & K_2O & & K_2O \\ x_{1.1} & + x_{2.1} & + \dots + & x_{12.1} \\ K_2O & K_2O & & K_2O \\ x_{1.2} & + x_{2.2} & + \dots + & x_{12.2} \\ \vdots & \vdots & \ddots & \vdots \\ K_2O & K_2O & & K_2O \\ x_{1.139} & + x_{2.139} & + \dots + & x_{12.139} \end{array} = \begin{array}{c} D_1^{K_2O} \\ D_2^{K_2O} \\ \vdots \\ D_{139}^{K_2O} \end{array} \right) \dots (6)''$$

$$x_{ij}^{K_2O} > 0 \quad (i=1,2,\dots,12, j=1,2,\dots,139) \dots (7)''$$

3.3.Evaluation of the Model(to be compared to the 1975)

If we compare the supposed deduction effectiveness by economic transportation L.P. model with record of 1975, we can expect shortening of transportation distance per ton and deduction effectiveness of transportation cost per ton.

3.3.1. Effects of Shortening Transportation Distance per T/T

3.3.1.1. Yearly Transportation Methods

By methods Yearly	current system	post system	Results	
			Km	%
Railway	272.4	165.3	107.1	39.3
Public road	32.7	23.6	9.1	27.8
Coast marine	398.5	242.0	156.5	39.3
Average	155.6	93.8	61.8	39.7

- (1). The average shortened transportation distance is 61.8 Km/Ton.(about 39.7% of the current distance)
- (2). The shortened railway is 107.1 Km/Ton.(about 39.3 % of the current distance)
- (3). The shortened public road is 9.1 Km/Ton.(about 27.8% of the current distance, the effects on shored distance between the warehouse are expected to be great)
- (4). The shortened coast marine transportation distance is 156.5 Km/Ton.(about 39.7% of the current distance)

3.3.1.2. Consumption, Transportation Methods

by consumption	Railway	Public road	Coast Marine	Average
Seoul, Gyeonggi	116.6	20.5	-	65.4
Gangwon	173.7	38.5	316.8	108.6
Chungbuk	164.3	17.8	-	86.0
Chungnam	160.0	20.9	-	88.1
Chunbuk	242.3	19.3	-	127.5
Chunnam	217.8	23.3	255.2	119.9
Gyeongbuk	141.5	19.2	171.3	80.6
Gyeongnam	86.8	31.6	103.7	56.1
Jaeju	-	44.8	269.5	157.2
Average	165.3	23.6	242.0	93.8

We know that railway transportation distance of Gangwon-Province is short to other area, but transportation distance of public road is long. We can know this is the reason why transportation destination is the mountain area of geographic condition.

3.3.2. Effects Saving Transportation Cost per Ton

- (1). The optimum transportation cost/Ton is 2,250.16

Classification		Quantity (M/T)	Amount (Won)	Unit Cost (Won)	Effects
Hypothesis	Optimum	1,940,710	4,366,899,080	2,250.16	2,423,000,000
	Current	1,940,710	6,790,544,000	3,499.00	
Results	Current	2,303,000	8,056,000,000	3,499.00	2,875,000,000
	Optimum	2,303,000	5,181,000,000	2,250.16	

Won. An annual saving transportation cost is expected at 2,423,000,000 Won (about 35.7% of the total transportation cost) based on the yearly total consumption is 1,941,710 M/T.

(2). An annual saving transportation cost is expected at 2,875,000,000 Won based on the 1975 which the total transportation size is 2,303,000 M/T.

We should classified to type for fertilizer, area, warehouse by need for deduction of transportation cost as above and we should deduce transportation cost to the respective transportation methods with consideration of special situation of transportation methods also or should harmonize it.

So we can consider the following deduction methods of transportation cost per ton.

First, we survey and estimate supply and demand of fertilizer to the respective area and after considering transportation capacity, we should select the optimal transportation network.

Second, we take premise for demand sufficiency to the respective transportation methods but we should peak demand point and selection of transportation methods should be taken after examination of profit and cost.

Third, we harmonise terminal to the respective transportation methods and modernize it. And we let it be convenient for freight working and custody.

Fourth, we construct circulation warehouse in the strategic site where is the concentrative demand area and should keep harmonization of transportation and custody

3.3.3. Geographical Assignment for Transportation of Fertilizer

3.3.3.1. Railway

See Fig. 6.

3.3.3.2. Public road

See Fig. 7.

3.3.3.3. Coast marine transportation

See Fig. 8.

4. SUGGESTION FOR IMPROVEMENT OF THE TRANSPORTATION SYSTEM

4.1. Suggestion for Development of a Total System² for Fertilizer

4.1.1. Process of the Model Analysis

System establishment of total supply and demand of fertilizer is approached by the four steps of analysis of the existing system, design a new system, evaluation of alternative system and final evaluation and we propose the basic direction of total management system which will progress as we analyze and examine the following matters on the basis of approached process to the respective steps.

(1). Study of the existing system

We collect the whole materials about fertilizer supply and demand for understanding the existing supply and demand management system of fertilizer in this step and enforce working process analysis through every kinds of surveys. As we establish the introductive frame of system to be fixed with basis of such a analysis, we should visualize sub-system for design a new system. Especially, distribution information to the existing transportation system should be understood in the establishing step of sub-system and at the same time, the other main establishment points of sub-system should be understood.

(2). Design a new system

We should confirm every kinds of survey matters which had been analyzed and examined in the first step with management indications of National Agricultural Cooperative Federation through information system at the same time and synthetic and systematic survey and analysis should be preceded to distribution and location of the respective area warehouse, capacity of transportation methods, transportation path of the existing respective transportation methods and various estimated ma-

2. For a Total Systems concept defined see, E.R.Dickey and N.L.Senesieb, "The Total System Concept", draft of entry for

terials for the reasonable study of the supply and demand forecast and stock of fertilizer, production plan and distribution system.

(3). Evaluation of Alternative System

Effective analysis should be achieved by system simulation which examines effectiveness, cost and items to new system.

(4). Final Evaluation

Main points as system enforcement to this will be made after we select the main alternative contents of alternative system. (See Fig. 9)

4.1.2. A New System of Total Distribution Model

New total distribution system of fertilizer is performed by report, co-ordination and information through the connective function of the respective information and transportation system as shown in Fig. 10.

4.2. Model for the Production and Demand Forecast by Kinds of Fertilizer

Model needed for production and demand forecast to the respective kind of fertilizer is Assignment Linear Programming Model to the respective elements and kinds of fertilizer to minimize the farmer delivery price which is constraints and objective function of the respective production by kinds and demand by elements of fertilizer. The composition is as follows.

$$\text{Min } Z = C_{a1.1}^1 X_1 + C_{a2.2}^2 X_2 + C_{a3.3}^3 X_3 + \dots + C_{an.n}^n X_n \quad \dots(8)$$

(C_{aj.}: Farmer Delivery Price + (Factory Price +

the Encyclopedia for Management, Reinhold Publishing Corp.
 "---Total System Concept: An approach to information systems design that conceives the business enterprise as an entity composed of interdependent systems and subystems, which, with the use of automatic data processing systems, attempts to provide timely and accurate management information which will permit optimum management decision making.---"

Average Transportation Cost)

X_{ij} : Supply to j th Demand Areas from i th Supply Areas (by kinds of fertilizer)

subject to

$$\left(\begin{array}{l} RD_{N.1}^{a1} x_1 + RD_{N.2}^{a2} x_2 + RD_{N.3}^{a3} x_3 + \dots + RD_{N.n}^{an} x_n = D_{N.}^n \\ SD_{P_{2O_5}.1}^{a1} x_1 + SD_{P_{2O_5}.2}^{a2} x_2 + SD_{P_{2O_5}.3}^{a3} x_3 + \dots + SD_{P_{2O_5}.n}^{an} x_n = D_{P_{2O_5}.}^n \\ TD_{K_{2O}.1}^{a1} x_1 + TD_{K_{2O}.2}^{a2} x_2 + TD_{K_{2O}.3}^{a3} x_3 + \dots + TD_{K_{2O}.n}^{an} x_n = D_{K_{2O}.}^n \end{array} \right) \dots (9)$$

$$\left(\begin{array}{l} x_1 \leq S_1^1 \\ x_2 \leq S_2^1 \\ x_3 \leq S_3^1 \\ \dots x_n \leq S_n^1 \end{array} \right) \dots (10)$$

$$x_j \geq 0 \quad (j=1,2,\dots,139) \quad \dots (11)$$

$$\left(\begin{array}{l} \sum_{j=1}^{139} RD_{Nj}^{an} x_j = D_{N.j}^n \quad (\text{Nitrogen}) \\ \sum_{j=1}^{139} SD_{P_{2O_5}j}^{an} x_j = D_{P_{2O_5}.j}^n \quad (\text{Phosphorus}) \\ \sum_{j=1}^{139} TD_{K_{2O}j}^{an} x_j = D_{K_{2O}.j}^n \quad (\text{Potash}) \end{array} \right) \begin{array}{l} \dots (9) \\ \text{Demand} \\ \text{Constraints} \end{array}$$

$$\left(\begin{array}{l} \sum_{i=1}^{12} x_{iN.}^{an} \leq S_{iN.}^N \quad (\text{Nitrogen}) \\ \sum_{i=1}^{12} x_{iP_{2O_5}.}^{an} \leq S_{iP_{2O_5}.}^{P_{2O_5}} \quad (\text{Phosphorus}) \\ \sum_{i=1}^{12} x_{iK_{2O}.}^{an} \leq S_{iK_{2O}.}^{K_{2O}} \quad (\text{Potash}) \end{array} \right) \begin{array}{l} \dots (10) \\ \text{Production} \\ \text{Constraints} \end{array}$$

$$(i=1,2,\dots,12)$$

$$\sum_{i=1}^{12} \sum_{j=1}^{139} C_{ij} x_{iN.}^{an} + \sum_{i=1}^{12} \sum_{j=1}^{139} C_{ij} x_{iP_{2O_5}.}^{an} + \sum_{i=1}^{12} \sum_{j=1}^{139} C_{ij} x_{iK_{2O}.}^{an} = \text{Min}(Z) \dots (8)$$

(Objective Function)

$$(i=1,2,\dots,12, j=1,2,\dots,139)$$

4.3. Expansion of the Model for Total Transportation System

Application of particular model of Linear Programming which is called to transportation L.P. model for establishing the total supply and demand management system by current of the optimal transportation of fertilizer for supply and consumption area.

That is to say, alternative variables to be considered from this study was X_{ij} be the supply to j th demand areas from i th supply areas, C_{ij} be average transportation cost to j th demand areas from i th supply areas (Factory price + Average transportation cost), S_i be supply capacity in i th supply areas and D_j be demand capacity in j th demand areas, and so the solution of transportation L.P. model is

$$\sum_{j=1}^m X_{ij} \leq S_i \quad \dots(12)$$

$$\sum_{i=1}^n X_{ij} \geq D_j \quad \dots(13)$$

$$X_{ij} \geq 0 \quad \dots(14)$$

and the model which should be applied hereafter as a model of fertilizer transportation to this develops it more and when consider $A_i \cdot SS_i$ be stock capability constraints i th supply area, $A_j \cdot DS_j$ be stock capacity j th demand area, $T_{ij} X_{ij}$ be transportation capacity to j th demand area from i th supply area, C_i be cost required stock i th supply area, C_j be stock cost j th demand area, $SS_{i,t-1}$ be former year stock of i th supply area, $DS_{j,t-1}$ be former year stock of j th demand area and TC_j be transportation capacity to j th demand area, the above solution will be

$$\sum_{j=1}^m X_{ij} + A_i \cdot SS_i \leq S_i + SS_{i,t-1} \quad \dots(15)$$

$$\sum_{i=1}^n X_{ij} + A_j \cdot DS_j = D_j + DS_{j,t-1} \quad \dots(16)$$

$$\sum_{i=1}^n T_{ij} X_{ij} \leq TC_j \geq D_j \quad \dots(17)$$

and minimize

$$\sum_{i=1}^n \sum_{j=1}^m C_{ij} X_{ij} + \sum_{i=1}^n C_i \cdot SS_i + \sum_{j=1}^m C_j \cdot DS_j \quad \dots(18)$$

Progress type of fertilizer transportation model considering such transportation capacity and custody capacity from the progress direction of future fertilizer transportation model as the above is adjusted as follows?

$$\text{Min}(Z) = \sum_{i=1}^n \sum_{j=1}^m C_{ij} X_{ij} + \sum_{j=1}^m C_{.j} \text{DS}_{.j} + \sum_{i=1}^n C_i \text{SS}_i \dots (19)$$

subject to

$$\begin{pmatrix} X_{1.1} + X_{1.2} + \dots + X_{1.n} + A_1 \text{SS}_1 \leq S_1 \\ X_{2.1} + X_{2.2} + \dots + X_{2.n} + A_2 \text{SS}_2 \leq S_2 \\ \vdots \\ X_{m.1} + X_{m.2} + \dots + X_{m.n} + A_m \text{SS}_m \leq S_m \end{pmatrix} \dots (15)$$

$$\begin{pmatrix} X_{1.1} + X_{2.1} + \dots + X_{m.1} + A_{.1} \text{DS}_{.1} = D_{.1} \\ X_{1.2} + X_{2.2} + \dots + X_{m.2} + A_{.2} \text{DS}_{.2} = D_{.2} \\ \vdots \\ X_{1.n} + X_{2.n} + \dots + X_{m.n} + A_{.n} \text{DS}_{.n} = D_{.n} \end{pmatrix} \dots (16)$$

$$\begin{pmatrix} T_{1.1} X_{1.1} + T_{2.1} X_{2.1} + \dots + T_{m.1} X_{m.1} \leq \text{TC}_{.1} \\ T_{1.2} X_{1.2} + T_{2.2} X_{2.2} + \dots + T_{m.2} X_{m.2} \leq \text{TC}_{.2} \\ \vdots \\ T_{1.n} X_{1.n} + T_{2.n} X_{2.n} + \dots + T_{m.n} X_{m.n} \leq \text{TC}_{.n} \end{pmatrix} \dots (17)$$

$$\begin{pmatrix} X_{ij} + A_i \text{SS}_i \leq S_i + \text{SS}_{i.t-1} & \text{Supply Constraints} \\ X_{ij} + A_{.j} \text{DS}_{.j} = D_{.j} + \text{DS}_{.j.t-1} & \text{Demand Constraints} \\ \sum_{j=1}^n T_{ij} X_{ij} \leq \text{TC}_{.j} \geq D_{.j} & \text{Transportation Capacity Constraints} \end{pmatrix} \dots (15''-17'')$$

And the established transportation L.P. model must enlarge the applied scope more and is more reasonable than any others. and the continual simulation to systemize the

3. The resulting model defines the so-called generalized transportation problem.

total supply and demand management should be kept pace with and the following matters should be considered on the established process of the future fertilizer supply and demand management system.

First, we considered the transportation cost including the incidental cost only as transportation cost between original and destination, but will include the productive cost price in the future study procedure and it is more reasonable and proper to make all the costs to the final destination for cost between original and destination.

Second, we must consider supply and demand quantities of original and destination and should establish the supply and demand quantities with all the warehouse construction plans and production planning to the respective area in the future.

Third, the optimal transportation plan should be made rather respective monthly and quarterly than yearly for low custody cost and reflecting the seasonal character of fertilizer transportation.

Fourth, transportation quantity coordinate in the transportation burden capacity scope is right if it is limitation to the transportation capacity between special original and destination. That is to say, transportation L.P. model can establish the most reasonable transportation distribution system to the transportation methods when the transportation demand between defined original and destination exceeds to transportation capability of the most profitable transportation methods.

Fifth, it is profit to divide the establishment of the final consumption area as possible as in detail on the total supply and demand management planning establishment of fertilizer. If the existing respective City and Kun classification divides into the respective unit warehouse, we can raise the transportation cost deduction effectiveness.

Sixth, as the existing transportation cost pool system is practicing with shortage of the realistic evident and scientific survey to the transportation record and total transportation amount which is the accounting basis of transportation cost freight unit price, the fertilizer transportation cost is possible to be increased at large and the budget is apt to be wasted. On the other hand, the freight unit price calculation of the transportation cost by transportation L.P. model (1) let the respective area's freight currency be the optimum (2) can calculate the minimum value or the optimum value of fertilizer transportation cost unit price per ton as to get the minimum value or the best proper value of the total transportation cost amount at the same time (3) if the transportation L.P. model is made, the re-

vision of alternative variables at any time is easy when the alterations of transportation cost unit price between original and destination and supply-demand quantities and we can get the solution of the economical transportation network and total amount of the minimum transportation cost correctly and fastly.

Seventh, we need to form fertilizer supply and demand zone to the respective factory and fertilizer elements for policy decision of warehouse facility enlargement in the optimal transportation system of fertilizer and economic transportation model conception and we must fix the O.R. model for location selection of warehouse. Location decision of warehouse dissolves not only bottle-neck of circulation path but also we should establish warehouse where we can keep and send to distribute the freight with minimum cost in the fastest time. And this forms a few sub-zone again from the warehouse enlargement to the respective supply and demand zone and we can define it as follows as a model to select the location at the proper district in the area.

Let W be the total cost of warehouse construction at jx site, $U_{f,jx}$ be the unit transportation cost from factory f to warehouse construction site jx , $X_{f,jx}$ be the total transportation cost from factory f to warehouse construction site, $U_{jx,j}$ be the unit transportation cost from warehouse construction site jx to demand area j , $X_{jx,j}$ be the warehouse construction cost in warehouse construction site j , f be location of factory(of fertilizer), j be demand area of fertilizer and jx be warehouse construction site of the several demand areas, then the O.R.model is given as follows.

$$\text{Min } W = \sum_f U_{f,jx} X_{f,jx} + \sum_j U_{jx,j} X_{jx,j} + C_{jx} \quad \dots (20)$$

And the circulation management to the respective warehouse by transportation cost minimization function model will get rid of inconvenience by distributing and sending the freight on the basis of the traditional administrative area unit.

APPENDIX

Table 1. Supply and Consumption Areas

Supply Areas		Consumption Areas	
		City-Province	City, Kun
Local Supplies	Choongju Fert.	Seoul	1
	Honam Fert.	Busan	1
	Youngnam Fert.	Gyeonggi	18
	Jinhae Fert.	Gangwon	15
	Kyeonggi Fert.	Chungbuk	10
	Poongnong Fert.	Chungnam	15
Import Ports	Busan	Chunbuk	13
	Yeosu	Chunnam	22
	Mokpo	Gyeongbuk	23
	Kunsan	Gyeongnam	19
	Changhang	Jaeju	2
	Incheon		
Total	12 Areas	139 Areas	

Note: Except for Ongjin-Kun, Gyeonggi-Province, Ulreung-Kun, Gyeongbuk-Province.

Table 2. Yearly Production (Unit: 1,000M/T)

Elements Yearly	Nitrogen		Phosphorus		Potash		Total	
	Q'ty	%	Q'ty	%	Q'ty	%	Q'ty	%
1969	367	65.3	146	26.0	49	8.7	562	100.0
1970	401	68.1	140	23.8	48	8.1	589	100.0
1971	408	67.5	146	24.2	50	8.3	604	100.0
1972	418	65.7	163	25.6	55	8.7	636	100.0
1973	447	65.7	160	23.6	73	10.7	681	100.0
1974	514	68.2	166	22.0	74	9.8	754	100.0
1975	583	66.6	196	22.4	97	11.0	876	100.0

Table 3. Consumption by Kinds of Fertilizer Unit: 1000M/T

Year Kinds	1971	1972	1973	1974	1975
Am. Sul.	8	1	-	25	74
Urea	519	559	572	704	767
Cal. Cyan.	18	23	19	4	-
Triple Sup. Phos.	24	20	12	179	193
Fused Phos.	149	190	350	141	129
Pot. Chlo.	42	51	96	124	152
Complex Fertilizer	542	580	705	601	570
Pot. Sul.	7	4	5	3	1
Others	-	-	-	-	55
Total	1,309	1,428	1,759	1,781	1,941

Table 4. Yearly Consumption Unit: 1000M/T

Year	1969	1970	1971	1972	1973	1974	1975
Q'ty	1,194	1,213	1,310	1,429	1,776	1,781	1,941
Increase Rate(%)	-	1.6	8.0	9.1	24.3	0.3	9.0

Table 5. Supply and Demand by Consumption Area
(in 1975) Unit: M/T

Element Area	Nitrogen(N)		Phosphorus(P ₂ O ₅)		Potash(K ₂ O)		Total	
	Q'ty	%	Q'ty	%	Q'ty	%	Q'ty	%
Seoul	5,181	59.3	2,350	26.9	1,204	13.8	8,735	100.0
Busan	1,453	62.0	582	24.9	307	13.1	2,342	100.0
Gyeonggi	117,579	55.4	61,934	29.2	32,720	15.4	212,232	100.0
Gangwon	59,676	55.2	32,231	29.8	16,306	15.0	108,233	100.0
Chungbuk	71,333	51.0	42,577	30.4	26,017	18.6	139,927	100.0
Chungnam	130,929	52.3	77,712	31.0	41,753	16.7	250,394	100.0
Chunbuk	126,988	53.4	71,478	30.0	39,493	16.6	237,959	100.0
Chunnam	186,248	54.6	103,696	30.4	51,069	15.0	341,013	100.0
Gyeongbuk	184,183	54.9	95,360	28.4	56,076	16.7	335,619	100.0
Gyeongnam	149,577	58.5	68,065	27.0	34,582	13.7	252,224	100.0
Jaeju	27,992	54.0	14,474	27.9	9,383	18.1	51,849	100.0
Total	1,061,158	55.7	570,459	30.0	308,910	16.2	1,940,527	100.0

Source; National Agricultural Cooperative Federation

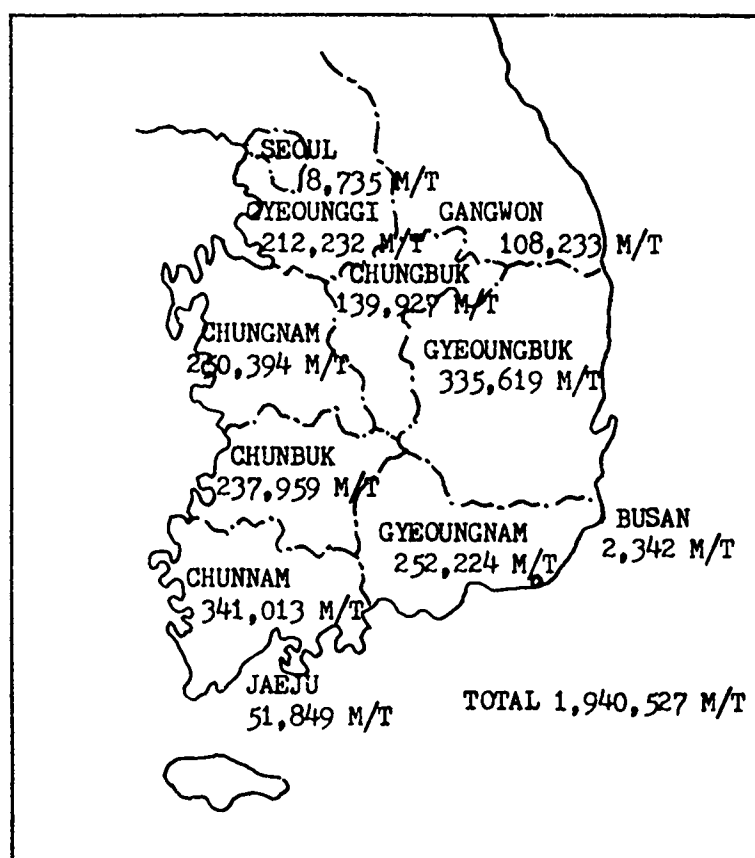


Fig. 1. Geographical Comparison of Supply and Demand by Elements of Fertilizer (in 1975)

Table 6. Yearly Transportation(M/T) Unit: 1000 M/T

Year	Railway		Public Road		Coast Marine		Total	
	M/T	%	M/T	%	M/T	%	M/T	%
1969	1,194	35.9	1,873	56.3	258	7.8	3,325	100.0
1970	1,051	34.4	1,815	59.5	185	6.1	3,051	100.0
1971	1,135	34.1	2,029	61.0	162	4.9	3,326	100.0
1972	1,259	39.6	1,847	58.1	75	2.3	3,181	100.0
1973	1,534	38.5	2,332	58.5	117	3.0	3,983	100.0
1974	1,878	44.5	2,242	53.1	105	2.4	4,224	100.0
1975	2,166	46.9	2,322	50.2	133	2.9	4,621	100.0

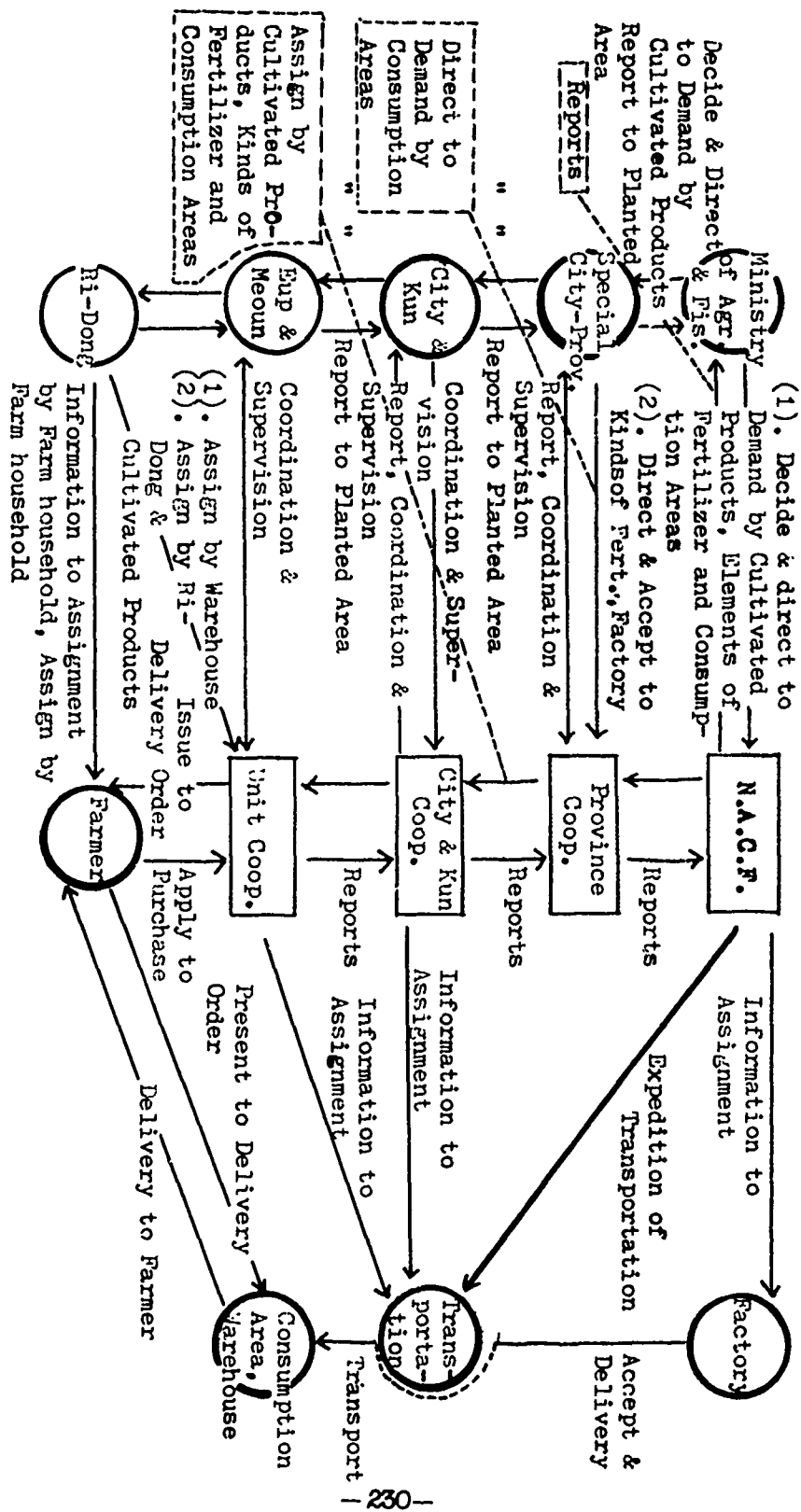


Fig. 2. Schematic Diagram for Supply of Fertilizer

Table 7. Yearly Transportation (M/T-Km) Unit:1,000M/T-Km

Year	Railway		Public Road		Coast Marine		Total	
	M/T-Km	%	M/T-Km	%	M/T-Km	%	M/T-Km	%
1969	330,198	71.5	46,825	10.1	84,743	18.4	461,766	100.0
1970	312,293	74.6	45,375	10.8	60,829	14.6	418,497	100.0
1971	344,399	74.2	65,417	14.3	53,450	11.5	464,266	100.0
1972	346,544	78.2	66,420	15.0	30,241	6.8	443,205	100.0
1973	419,563	79.6	60,464	11.5	46,796	8.9	526,823	100.0
1974	524,915	81.6	76,315	11.9	42,091	6.5	643,321	100.0
1975	590,000	82.1	76,000	10.6	53,000	7.3	719,000	100.0

Table 8. Caculative applied cost for fertilizer transportation unit price to respective transportation methods

Methods	Freight Charge	Incidental Cost	Remark
Railway	Basic Freight Charge, Vacant Vehicle Return Charge	Loading and Unloading Charge, Damaged Charge, Packing Charge, Inspection Charge, Dispatch Charge to Factory	Except for Custody Charge, Removal Charge
Public-road	Basic Freight Charge	Incidental Cost of Railway is applied	Except for Custody Charge, Removal Charge, In and Out Warehouse Charge and Special Discount Charge
Coast Marine	Basic Freight Charge	Loading and Unloading Charge	" (Fertilizer is accounted by 10% discount Increase charge)

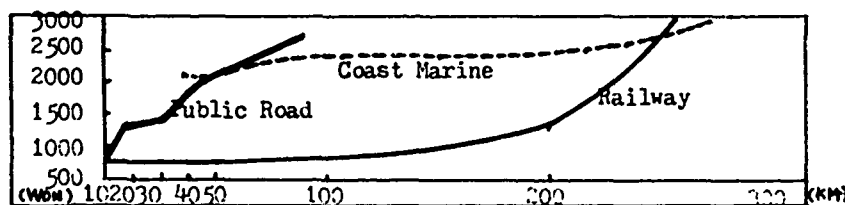


Fig. 3. Comparison of Transportation Distance-Cost by Methods of Fertilizer(Per M/T)

Table 9. Comparison of Transportation Cost
Per M/T-Km by Methods (by Items) Unit: Won per M/T

Methods Items	Railway		Public Road		Coast Marine	
	Amount	%	Amount	%	Amount	%
Total Cost	2,303.03	100.0	2,291.32	100.0	2,348.72	100.0
Freight Charge	1,834.80	79.7	1,860.83	81.2	1,171.72	49.9
Basic	1,668.00	72.4	1,860.83	81.2	1,171.72	49.9
Others	166.80	7.3	-	-	-	-
Incidental Cost	468.23	20.3	430.49	18.8	1,177.00	50.1
Loading	137.83	6.0	137.83	6.0	371.25	15.8
Unloading	125.29	5.4	125.29	5.5	371.25	15.8
Damaged & Packing	162.36	7.1	162.36	7.1	434.50	18.5
Inspection	5.01	0.2	5.01	0.2	-	-
Others	37.74	1.6	-	-	-	-

Note;

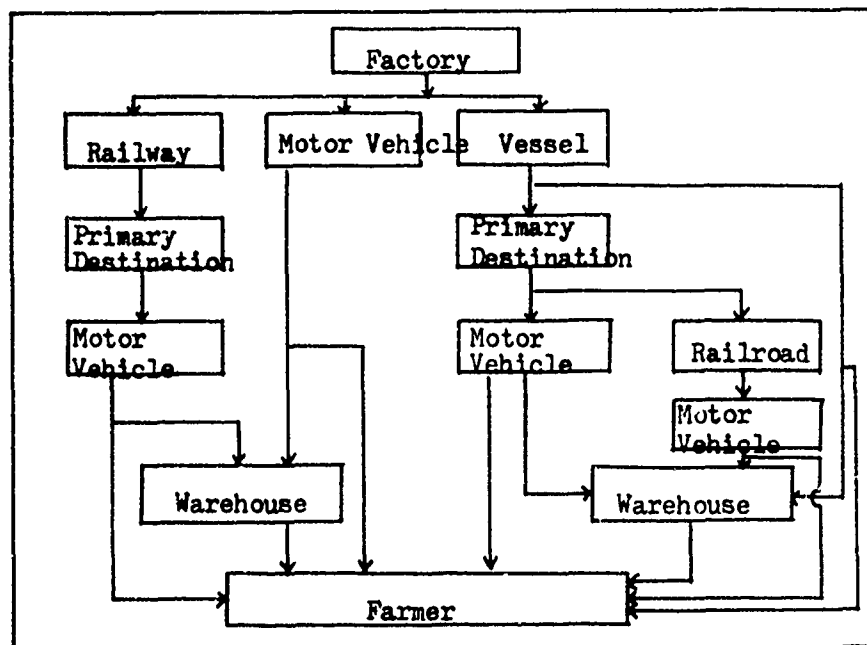


Fig. 4. Patterns of Transportation Path by Methods

Table.10. Illustration for the Primary Transportation Relay Place by Originals

Demand	Supply	Domestic Supply Areas						Import Ports					
		Choongju	Honam	Youngnam	Jinhae	Gyeonggi	Poong	Busan	Yeosu	Mokpo	Kunsan	Changhang	Incheon
Gyeonggi-pro	Choongju	Choongju	Choongju	Choongju	Choongju	Choongju	Choongju	Choongju	Choongju	Choongju	Choongju	Choongju	Choongju
Yangju-kun	Choongju	Choongju	Choongju	Choongju	Choongju	Choongju	Choongju	Choongju	Choongju	Choongju	Choongju	Choongju	Choongju
Gangwon-pro.	Jaechun	Jaechun	Jaechun	Jaechun	Jaechun	Jaechun	Jaechun	Jaechun	Jaechun	Jaechun	Jaechun	Jaechun	Jaechun
Youngwoel-kun	Jaechun	Jaechun	Jaechun	Jaechun	Jaechun	Jaechun	Jaechun	Jaechun	Jaechun	Jaechun	Jaechun	Jaechun	Jaechun
Chungbuk-pro.	Jung	Okeun	Okeun	Okeun	Okeun	Okeun	Okeun	Okeun	Okeun	Okeun	Okeun	Okeun	Okeun
Jinchun-kun	Jeong	Jeong	Jeong	Jeong	Jeong	Jeong	Jeong	Jeong	Jeong	Jeong	Jeong	Jeong	Jeong
Chunnam-pro.	Mokpo	Mokpo	Mokpo	Mokpo	Mokpo	Mokpo	Mokpo	Mokpo	Mokpo	Mokpo	Mokpo	Mokpo	Mokpo
Shinan-kun	Mokpo	Mokpo	Mokpo	Mokpo	Mokpo	Mokpo	Mokpo	Mokpo	Mokpo	Mokpo	Mokpo	Mokpo	Mokpo
Gyeongnam-pro.	Kunbuk	Kunbuk	Kunbuk	Kunbuk	Kunbuk	Kunbuk	Kunbuk	Kunbuk	Kunbuk	Kunbuk	Kunbuk	Kunbuk	Kunbuk
Hapchun-kun	Kunbuk	Kunbuk	Kunbuk	Kunbuk	Kunbuk	Kunbuk	Kunbuk	Kunbuk	Kunbuk	Kunbuk	Kunbuk	Kunbuk	Kunbuk
Jaeju-pro.	Kunsan	Mokpo	Busan	Jinhae	Incheon	Kunsan	Kunsan	Jaeju	Jaeju	Jaeju	Jaeju	Jaeju	Jaeju
Bukjaeju-kun	Jaeju	Jaeju	Jaeju	Jaeju	Jaeju	Jaeju	Jaeju	Jaeju	Jaeju	Jaeju	Jaeju	Jaeju	Jaeju

Table 11. Illustration for the Transportation Unit Cost by Originals Unit: Won

Demand	Supply	Domestic Supply Areas						Import Ports					
		Choongju	Honam	Youngnam	Jinhae	Gyeonggi	Poong	Busan	Yeosu	Mokpo	Kunsan	Changhang	Incheon
Yangju-kun	Choongju	175309	236769	267349	267349	129739	206189	267349	267349	252059	206189	145029	145029
Youngwoel-kun	Choongju	171871	278901	248321	278901	217741	263611	263611	263611	263611	248321	233031	233031
Jinchun-kun	Choongju	218760	279920	295210	295210	249340	264630	295210	295210	295210	249340	249340	249340
Shinan-kun	Choongju	261990	170250	307860	261990	277280	292570	352572	352572	352572	352572	352572	352572
Hapchun-kun	Choongju	452186	391026	375726	345156	467426	482766	360446	360446	406316	406316	457476	457476
Bukjaeju-kun	Choongju	536067	455547	474707	412794	538487	566447	412794	412794	378144	412794	412794	476374

Table 12. Supply Capacity by Supply Areas

Unit: 1000 H/F

Supply Areas	Elements S_i	Nitrogen(N)		Phosphorus(P_2O_5)		Potash(K_2O)	
		S_i^N	Supply Capacity	$S_i^{P_2O_5}$	Supply Capacity	$S_i^{K_2O}$	Supply Capacity
Choongju	S_1	S_1^N	310,252	$S_1^{P_2O_5}$	-	$S_1^{K_2O}$	-
Honam	S_2	S_2^N	133,648	$S_2^{P_2O_5}$	-	$S_2^{K_2O}$	-
Youngnam	S_3	S_3^N	707,513	$S_3^{P_2O_5}$	161,712	$S_3^{K_2O}$	95,974
Jinhae	S_4	S_4^N	183,764	$S_4^{P_2O_5}$	127,151	$S_4^{K_2O}$	73,954
Gyeonggi	S_5	S_5^N	-	$S_5^{P_2O_5}$	144,621	$S_5^{K_2O}$	-
Poongnong	S_6	S_6^N	-	$S_6^{P_2O_5}$	135,924	$S_6^{K_2O}$	-
Busan	S_7	S_7^N	-	$S_7^{P_2O_5}$	72,275	$S_7^{K_2O}$	91,386
Yeosu	S_8	S_8^N	-	$S_8^{P_2O_5}$	31,238	$S_8^{K_2O}$	15,413
Hokpo	S_9	S_9^N	-	$S_9^{P_2O_5}$	14,648	$S_9^{K_2O}$	50,605
Kunsan	S_{10}	S_{10}^N	6,667	$S_{10}^{P_2O_5}$	6,075	$S_{10}^{K_2O}$	36,066
Changchang	S_{11}	S_{11}^N	-	$S_{11}^{P_2O_5}$	20,121	$S_{11}^{K_2O}$	17,177
Incheon	S_{12}	S_{12}^N	14,307	$S_{12}^{P_2O_5}$	47,328	$S_{12}^{K_2O}$	58,680
Total	$\sum_{i=1}^{12} S_i$	$\sum_{i=1}^{12} S_i^N$	1,356,151	$\sum_{i=1}^{12} S_i^{P_2O_5}$	761,093	$\sum_{i=1}^{12} S_i^{K_2O}$	439,175

Table 13. Consumption by Elements, Demand Areas
for Fertilizer Unit: 1000 M/T

Elements		Nitrogen (N)		Phosphorus (P ₂ O ₅)		Potash (K ₂ O)	
Demand Areas	D _j	D _j ^N	Q'ty	D _j ^{P₂O₅}	Q'ty	D _j ^{K₂O}	Q'ty
Seoul	D ₁	D ₁ ^N		D ₁ ^{P₂O₅}		D ₁ ^{K₂O}	
Incheon	D ₂	D ₂ ^N		D ₂ ^{P₂O₅}		D ₂ ^{K₂O}	
Yangju-Kun	D ₃	D ₃ ^N		D ₃ ^{P₂O₅}		D ₃ ^{K₂O}	
Yeoju-Kun	D ₄	D ₄ ^N		D ₄ ^{P₂O₅}		D ₄ ^{K₂O}	
Pyeongtak-Kun	D ₅	D ₅ ^N		D ₅ ^{P₂O₅}		D ₅ ^{K₂O}	
⋮	⋮	⋮		⋮		⋮	
Bukjaeju-Kun	D ₁₃₈	D ₁₃₈ ^N		D ₁₃₈ ^{P₂O₅}		D ₁₃₈ ^{K₂O}	
Namjaeju-Kun	D ₁₃₉	D ₁₃₉ ^N		D ₁₃₉ ^{P₂O₅}		D ₁₃₉ ^{K₂O}	
Total	$\sum_{j=1}^{139} D_j$	$\sum_{j=1}^{139} D_j^N$		$\sum_{j=1}^{139} D_j^{P_2O_5}$		$\sum_{j=1}^{139} D_j^{K_2O}$	

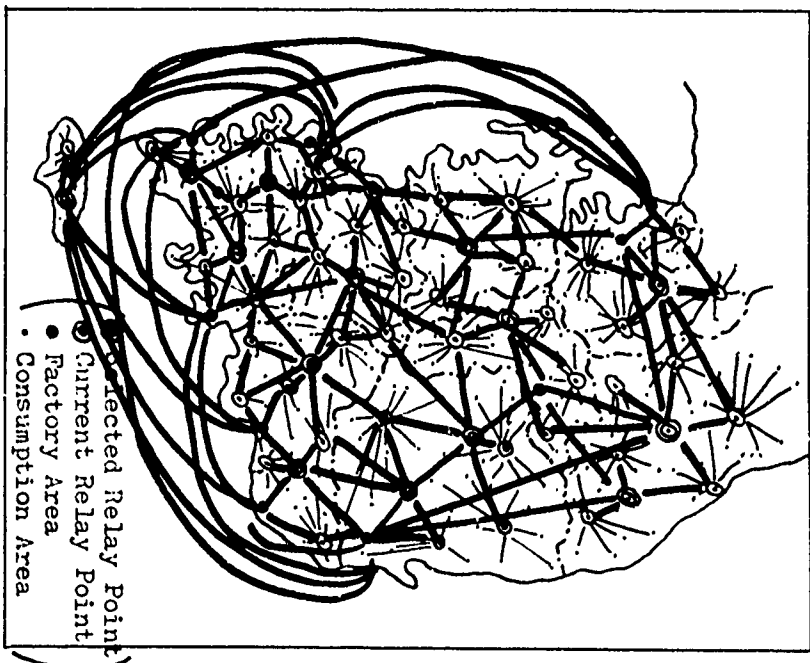


Fig. 5. Illustration for the Selection of Relay Points

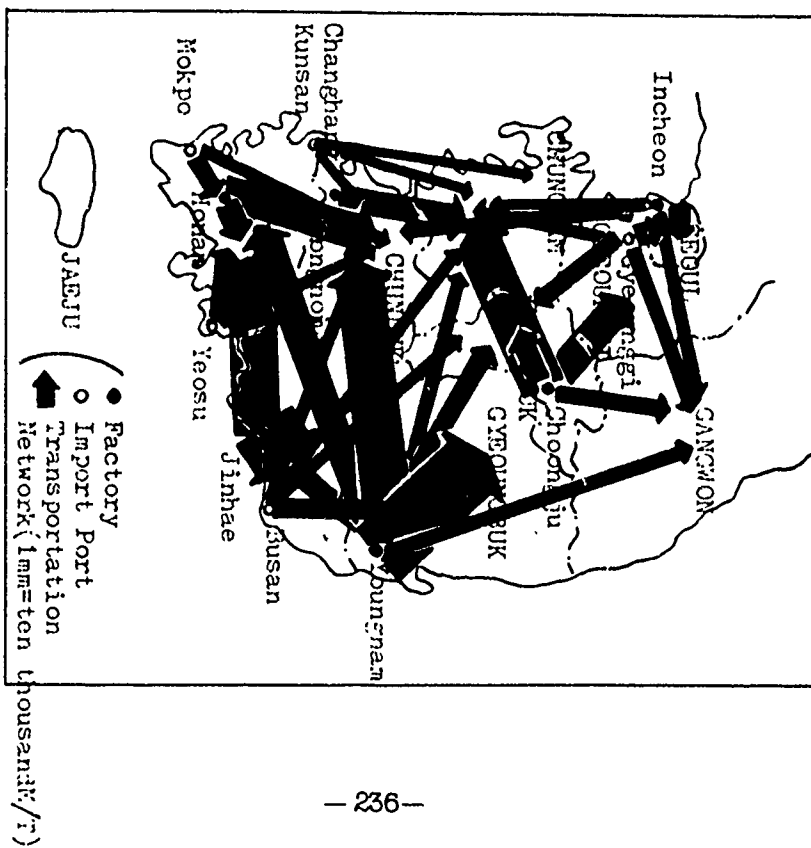


Fig. 6. Geographical Assignment for Transportation of Fertilizer by Railway

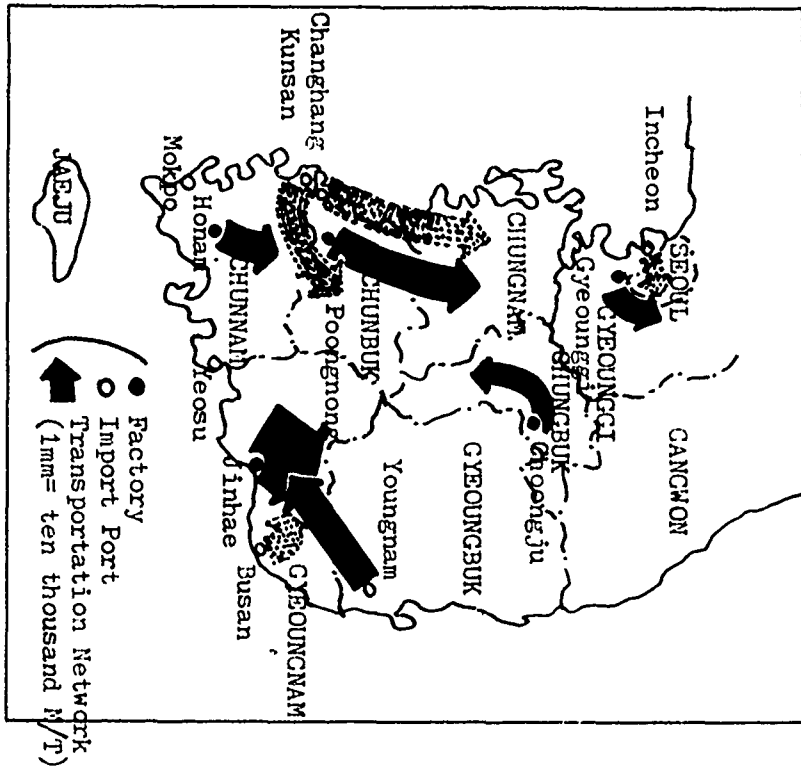


Fig. 7. Geographical Assignment for Transportation of Fertilizer by Public Road

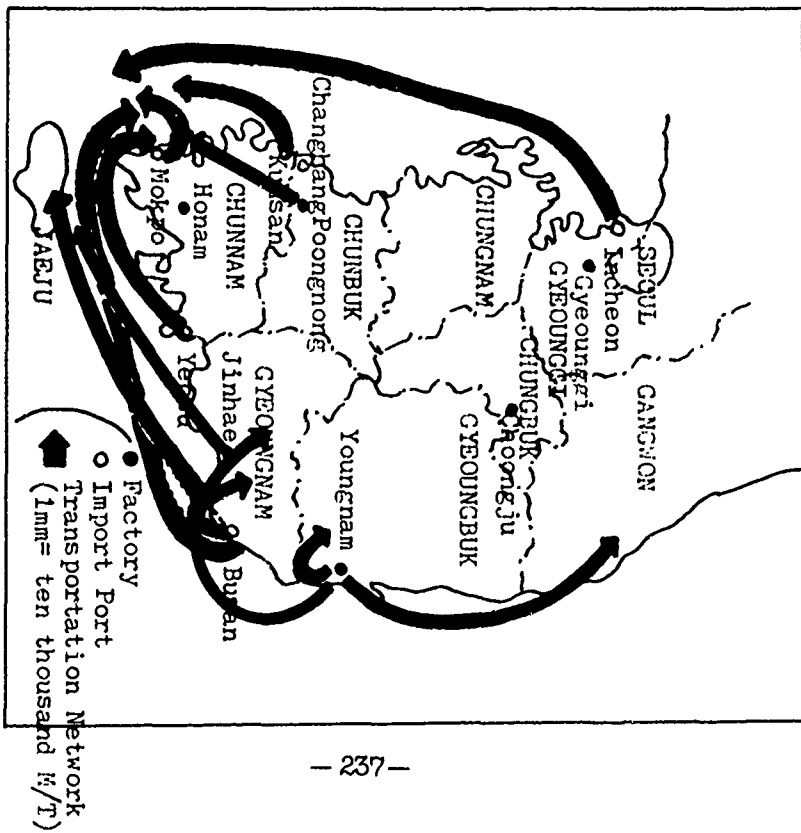


Fig. 8. Geographical Assignment for Transportation of Fertilizer by Coast Marine

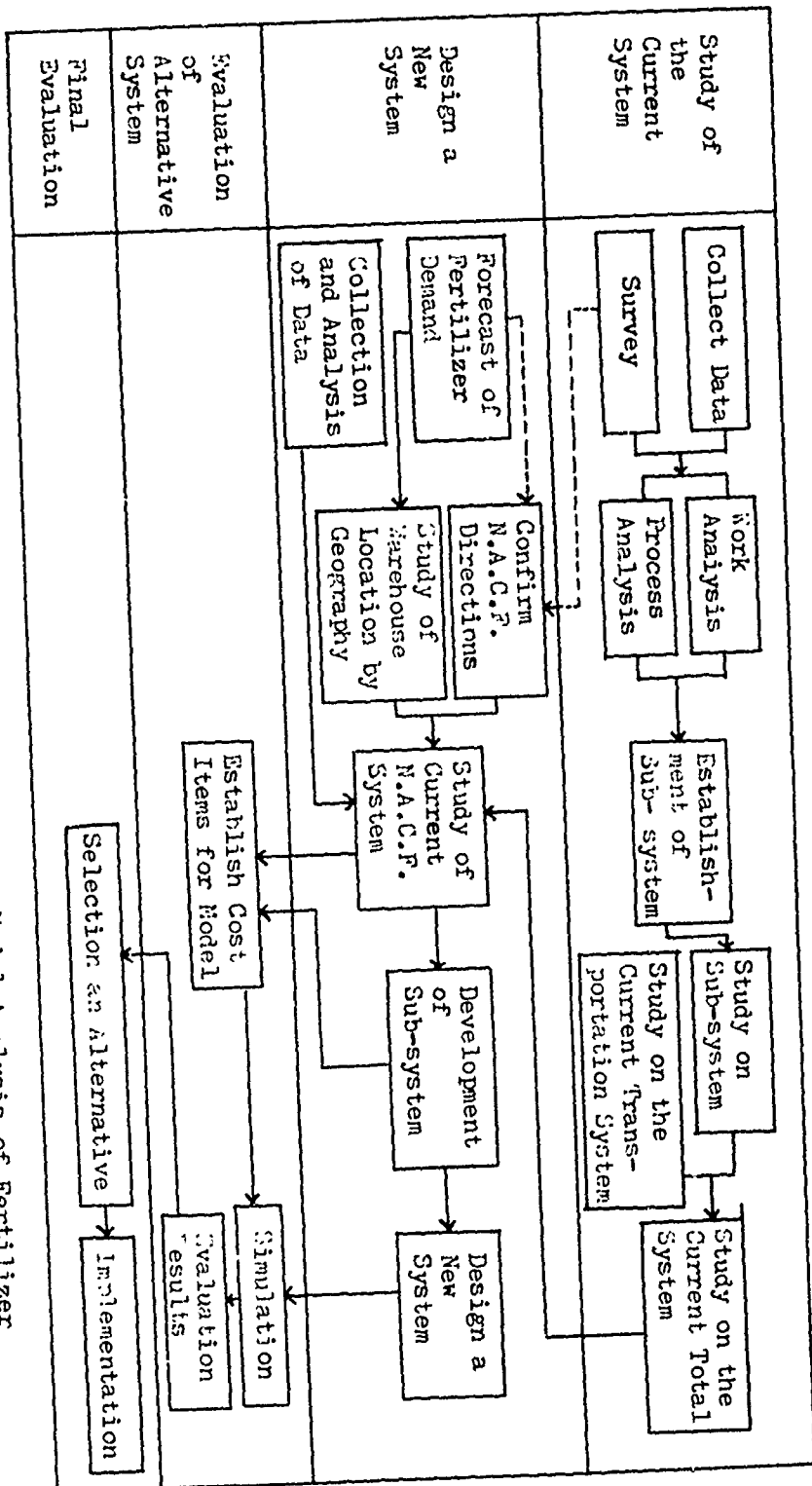


Fig. 9. Flow of Process for the Total System Model Analysis of Fertilizer

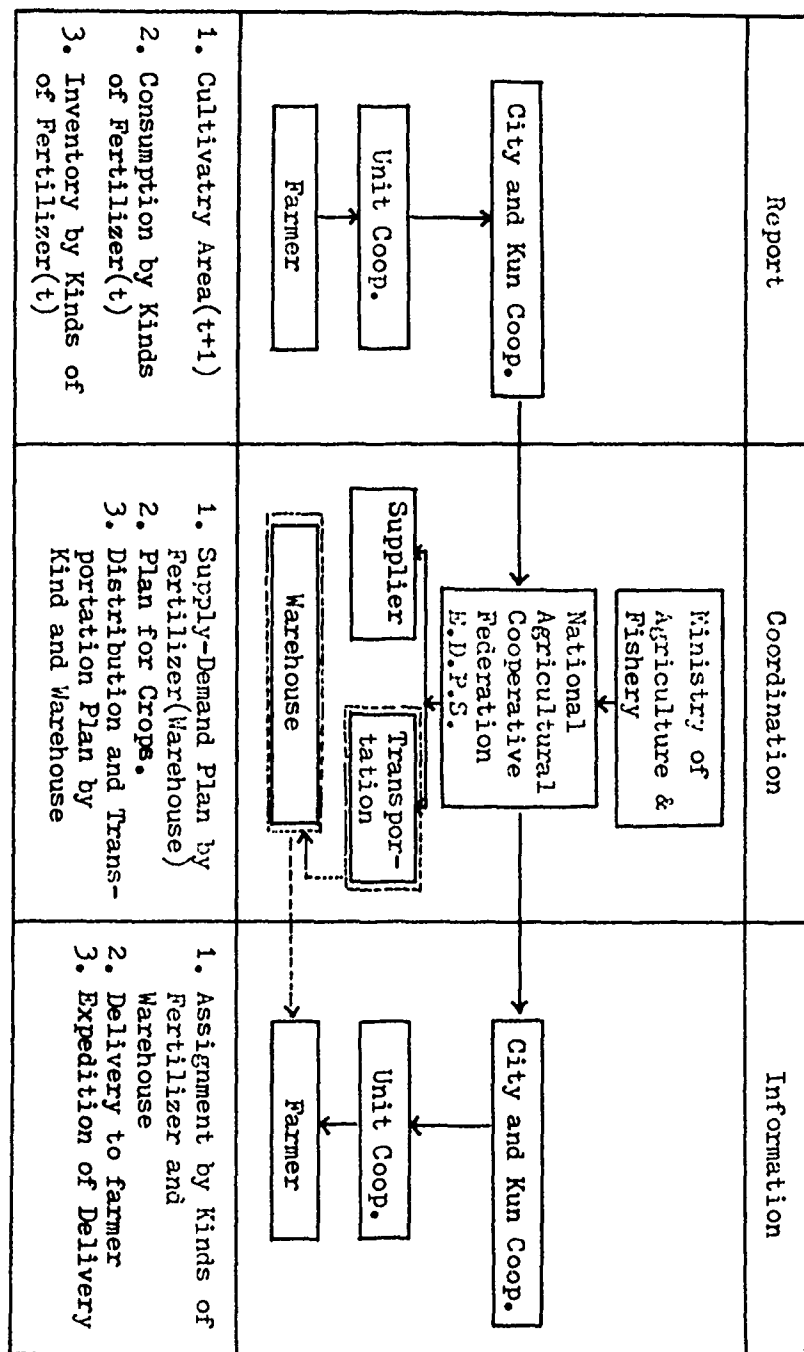


Fig. 10. Flow of a New System of Total Distribution Model

AN ALTERNATIVE ZERO-ONE OPTIMIZATION MODEL FOR THE LOCATION OF FIRE STATIONS

D. VAN OUDHEUSDEN and F. PLASTRIA

Centrum voor Statistiek en Operationeel
Onderzoek
Vrije Universiteit Brussel
Pleinlaan 2, Brussels, Belgium

ABSTRACT. The most important feature of locating fire stations is the possibility of reaching every point in a given area within a fixed time. As a completely sure location pattern cannot be realised in our traffic congested cities, we propose a model in which, given the number of fire stations, the probability of reaching any point within the statutory time is maximized. If necessary, this probability can be increased for certain zones of the city. The model takes into account that at least two fire stations will send fire engines to the place of disaster. Two implicit enumeration algorithms are proposed for the resolution of the optimization model. Some computation results, showing the feasibility of the procedure, are given.

1. INTRODUCTION

In a recent study J.N.M. van Loon and J.A.M. Schreuder [1] treat the location of fire stations in the city of Rotterdam, Holland. The study is certainly one of the most realistic ever made and reviews the specific requirements encountered in this particular field. It makes clear that, for the location purposes, it is most important to be sure to be in a position to reach any point in the urban area within a fixed time (6 or 8 min.). As a matter of fact, fire extinction is only possible before the self enlightening temperature of the materials has been reached and is therefore often a question of minutes.

J.N.M. van Loon and J.A.M. Schreuder convert the problem of determining the minimum number and location of fire stations into a set covering problem. First of all, they select, according to well defined rules, a finite set S of sites in which fire stations can be located. They then calculate all average travel times $d(s_i, t_j)$ for a fire engine going from $s_i \in S$ to t_j , the point in the urban district T_j least accessible from s_i . With these values $d(s_i, t_j)$, the formulation of the set covering problem is quite obvious. One has to select a minimal set $R \subset S$ of fire stations so that every T_j is "covered" within the statutory time: for every T_j a fire station $s_k \in R$ must exist such that $d(s_k, t_j) \leq 6 \text{ min.}$

J.N.M. van Loon and J.A.M. Schreuder require a "double covering", instead of a single one, as usually at least two fire stations will send fire engines to the place of disaster. The reason for alerting at least two fire stations is the fear of being blocked by the traffic, this is not unlikely for example during rush hours. It is this important aspect which suggests a probabilistic approach.

Therefore we consider probabilities $P(s_i, t_j) = p_{ij}$ of arriving within the statutory time instead of the travel times $d(s_i, t_j)$. We then construct a location model that, given the number of fire stations, minimizes the probability v that no fire engine will arrive within the necessary time in the least accessible district.

$$\min v \quad (1)$$

$$\prod_i (1 - p_{ij} y_i) \leq v \quad \forall j \quad (2)$$

$$\sum_i y_i = p \quad (3)$$

$$y_i \in B \stackrel{N}{=} \{0, 1\} \quad \forall i \quad (4)$$

y_i is equal to 1 if a fire station is established in s_i and equal to 0 otherwise. p is the fixed number of fire stations.

One can note that in the model, every fire station having a non-zero probability of reaching the considered district T_j in time, is supposed to be alerted. In practice, the number of fire stations with a non-zero probability will very often be only two or three.

It is easy to show that (1) - (4) is equivalent to the following linear optimization model:

$$\min (\ln v) \quad (5)$$

$$\sum_i y_i \ln (1-p_{ij}) \leq \ln v \quad \forall j \quad (6)$$

$$\sum_i y_i = p \quad (7)$$

$$y_i \in B \quad \forall i \quad (8)$$

It is the resolution of this optimization model that the paper deals with.

It may be important to note that

- It is possible to require a lower probability of not arriving in time for certain districts. In that case the corresponding p_{ij} -values have to be multiplied by an appropriate factor.
- The p_{ij} values can vary considerably during each 24 hour period. The model can first be solved for the smallest p_{ij} values (for example those corresponding to rush hours) and with the location obtained in this way one can subsequently examine whether some of the fire stations can be closed during calmer periods.

2. MATHEMATICAL NOTATIONS AND SOME RESULTS

Name I a finite set with cardinal number n and $g: I \rightarrow R$ a real valued function defined on I , we will write $\sum_I g$ for $\sum_{i \in I} g(i)$ and $\min_I g$ for $\min \{g(i) | i \in I\}$. If J is a subset of I we will use $\sum_J g$ and $\min_J g$ instead of the rather cumbersome notation $g|_J$ for the restriction of g to J .

The problem concerned can now easily be stated as follows:

Given the set of functions $f_k: I \rightarrow R$ ($k \in K$), find $J^* \subset I$ such that

$$\min_K \sum_J f_k = \max \{ \min_K \sum_J f_k \mid J \subset I \text{ and } |J|=p \} \quad (9)$$

Without loss of generality we may suppose all functions f_k to be positively valued, since by adding a constant c to all functions f_k ($k \in K$) we find a new problem of type (9)

which generates the same optimal solution J^* and adds p_c to the optimal value of the objective function.

It is also clear that if some function f_k dominates other functions f_e , i.e. $f_k(i) \geq f_e(i)$ for all $i \in I$, then, if we consider problem (9) with the function f_e left out, we still obtain the same optimal solution J^* and keep the same optimal value for the objective function.

Since we want to construct branch and bound algorithms, we have to be able to determine an upper bound for the solution to the reduced problem. A reduced problem is obtained from the original problem by selecting some elements of I , refusing others, and leaving all others free.

Allowing J_0 to be the set of refused elements of I , J_1 the set of selected elements of I , and J_2 the set of free elements of I , and J_2 the set of free elements of I . The reduced problem is finding:

$$a(J_0, J_1) = \max \{ \min_K \sum_{J_1 \cup J_k} f_k \mid J \subseteq J_2 \text{ and } |J| = p-s \} \quad (10)$$

where s stands for the cardinal number of J_1 .

If we want this reduced problem to have any sense, we must have $s = |J_1| < p$ and $t = |J_0| < n-p$.

For each $k \in K$ we can split J_2 in two subsets J_k^+ and J_k^- such that $|J_k^+| = p-s$ and

$$S_k \stackrel{N}{=} \min_{J_k^+} f_k \geq \max_{J_k^-} f_k \stackrel{N}{=} s_k.$$

Then call $M_k \stackrel{N}{=} \sum_{J_k^+ \cup J_1} f_k$ and $M(J_0, J_1) \stackrel{N}{=} \min_K M_k$.

The values S_k , s_k and M_k are well defined, even if the sets J_k^+ and J_k^- are not in the special case where $S_k = s_k$.

It is easy to show that $M(J_0, J_1) \geq a(J_0, J_1)$ and so it is an upper bound for the problem (10).

The efficiency of a branch and bound algorithm can be improved by the use of penalties [2], [3]. These are amounts calculated for each element of J_2 (free element of I) which indicate at least how much the upper bound of (10) will decrease by adding the free element to J_0 (refusing + down penalty) or to J_1 (selecting + up penalty).

For each $i \in J_2$ call

$$p_i^k = \begin{cases} 0 & \text{if } f_k(i) \leq s_k \\ M(J_0, J_1) - M_k + f_k(i) - s_k & \text{if } f_k(i) > s_k \end{cases} \quad (11)$$

and

$$q_i^k = \begin{cases} 0 & \text{if } f_k(i) \geq s_k \\ M(J_0, J_1) - M_k - f_k(i) + s_k & \text{if } f_k(i) < s_k \end{cases} \quad (12)$$

Then $P_i^0 = \max_k q_i^k$ and $P_i^1 = \max_k q_i^k$ represent such penalties.

It is indeed easy to show that

$$M(J_0 \cup \{i\}, J_1) \leq M(J_0, J_1) - P_i^0 \quad (13)$$

and

$$M(J_0, J_1 \cup \{i\}) \leq M(J_0, J_1) - P_i^1 \quad (14)$$

3. ALGORITHMS

We have developed two algorithms for solving problem (9). The first one is a simple lexicographic implicit enumeration technique based on the upper bound found above. The second one makes use of the penalties defined above.

3.1. Algorithm 1:

To improve the efficiency of a lexicographic enumeration technique it is important to have an a priori idea of the elements of I which probably will be chosen, and base the order in which the elements of I will be fixed on this probability. For each element i of I we consider the values $f_k(i)$ ($k \in K$). If the mean value m_i of the $f_k(i)$ is high, and the variation s_i of these values is low, the element i of I has a good chance of being chosen. So the values m_i/s_i can be viewed as a measure of the probability of i to be chosen.

Thus the algorithm starts by ordering the elements of I in function of decreasing values of m_i/s_i . Choosing the first p elements of I according to this order gives a first solution to the problem (9), which can be viewed as a heuristic solution giving, in other algorithms, a starting value for the objective function.

The algorithm now is as follows.

(1) The order of I and initial values.

Arrange I according to decreasing values of m_i/s_i , calculate the heuristic solution, i.e. J_1 gets the p first elements of I , Z the value of the objective function, $J_0 = \emptyset$, $J_2 = I \setminus J_1$. Go to (4).

(2) Forward step.

Add the first free element of I to J_1 ; if there are no further free elements ($J_2 = \emptyset$) go to (4).

If J_1 now has p elements, calculate the value of the objective function for this solution and compare with the current

best solution. Go to (4).

(3) Test.

Calculate $M(J_0, J_1)$. If $M(J_0, J_1) \geq Z$ go to (2).

(4) Backward step.

If the last fixed element belongs to J_1 , transfer this to J_0 and go to (2). If not, free this element and go to (4). If there are no more fixed elements the algorithm terminates.

3.2. Algorithm 2:

This algorithm uses the penalties as defined above for lowering the upper bound of the reduced problem, and helping to choose the free element of I the value of which has to be fixed in a forward step. We also introduce conditional tests which can greatly reduce the dimension of the solution tree. In the following algorithm J represents the current best solution, and Z the current best value for the objective function.

(1) Initial values.

$J_0 = \emptyset$, $J_1 = \emptyset$, $J_2 = I$, $Z = 0$ (or the value found with any heuristic), $J = \emptyset^2$ (or the solution found with this heuristic).

(2) Resolution test.

- a) If $|J_0| = n - p$, set $J_1 = J_1 \cup J_2$, go to c)
- b) If $|J_1| \neq p$ go to (3)
- c) Calculate $a = \min_k \sum_{J_1} f_k$

If $a > Z$, set $Z = a$ and $J = J_1$

In any case go to (7).

(3) First optimality test.

Find $M(J_0, J_1)$. If $M(J_0, J_1) < Z$ go to (7).

(4) Second optimality test.

For each $i \in J_2$ find $v_i = M(J_0, J_1) - \min(P_i^0, P_i^1)$. If any $v_i < Z$ go to (7).

(5) Conditional optimality test.

For each $i \in J_2$ find $w_i = v_i - |P_i^0 - P_i^1|$

If $w_i < Z$ and

- a) $P_i^0 > P_i^1$ and

- If $|J_1| < p$, set $J_1 = J_1 \cup \{i\}$
 $J_2 = J_2 \setminus \{i\}$

- If $|J_1| = p$, go to (7).

- b) $P_i^0 \leq P_i^1$ and

- If $|J_0| < n - p$, set $J_0 = J_0 \cup \{i\}$
 $J_2 = J_2 \setminus \{i\}$

- If $|J_0| = n - p$, go to (7).

If in this step at least one element i has been fixed, go to (2).

(6) Choice.

Find a $j \in J_2$ with maximum difference between J_0 up and down penalty. Set $J_2 = J_2 \setminus \{j\}$ and $J_1 = J_1 \cup \{j\}$ if $P_j^0 > P_j^1$, or $J_0 = J_0 \cup \{j\}$ if $P_j^0 \leq P_j^1$. Go to (2).

(7) Backtracking.

Find the last element j which has been fixed in step (6). If no such element can be found the algorithm terminates. Otherwise transfer this j from J_0 to J_1 or from J_1 to J_0 whichever is correct and stop considering j as fixed at step (6). Cancel in J_0 and J_1 and transfer to J all elements fixed in steps (5) and (7), after the fixing of j in step (6). Go to (2).

4. COMPUTATIONAL EXPERIENCE

Both algorithms (AL1 and AL2) were programmed in the code FORTRAN IV ext. and processed on a CDC 6500. The coefficients $f_k(i)$ are randomly generated integers between 0 and 100. As we consider calculation times too dependent on programmer and machine, there was not much attention given to finding average times for different problem sizes.

Table 1 gives an idea of the results for $|K| = 5$ and $|I| = 30$.

Table I

p	AL1 (20 probl.)			AL2 (5 probl.)		
	min	avg	max	min	avg	max
3	.319	.507	.751	.326	.341	.366
4	1.122	2.025	3.099	.441	.497	.629
5	2.816	6.941	12.875	.434	1.174	2.486

The above table clearly indicates the superiority of AL2 although no heuristic preceded it. Both algorithms will surely profit by introducing a heuristic solution before proceeding.

More attention was given to the evolution of the calculation times with changing $|K| = m$, $|I| = n$, p , always keeping both other parameters constant.

Table 2 gives the calculation times found when $n = 20$ and $p = 5$ for m varying from 5 up to 35. These results show that, as expected, the upper bound used in both algorithms gets less efficient as m increases. Obviously the

increase of the number of districts to get more precise probabilities, is feasible.

Table 3 shows the evolution for n varying from 10 to 100, with $m = 3$ and $p = 5$. It is clear from this table that the number of possible locations may be drastically increased.

Table 4 gives an idea of the calculation times for changing p . As p increases to $n/2$, calculation times increase enormously. The same happens when p decreases from n to $n/2$ which was to be expected.

It can therefore be concluded from these results that problems of small to medium sizes can be solved by AL2. Most restricted of all is the number of fire stations to be located. As in reality this number is rather small (not more than 10 in [1]), the proposed optimization model and algorithm may be useful for real problems of fire station location.

Table 2

$n = 20 \quad p = 5$		
m	AL1	AL2
5	1.910	.910
7	1.765	.732
9	>2	1.259
11	-	1.667
13	-	1.801
15	-	1.949
17	-	4.095
19	-	3.353
21	-	4.229
23	-	6.481
25	-	7.559
27	-	4.843
29	-	7.164
31	-	9.449
33	-	9.539
35	-	14.900

Table 4

$m = 5 \quad n = 40$		
p	AL1	AL2
2	.237	.499
3	1.201	.746
4	5.763	1.116
5	32.403	5.302
6	>30	9.838
7	-	32.802
8	-	58.026
9	-	>100

Table 3

$m = 3 \quad p = 5$		
n	AL1	AL2
10	.064	.034
13	.101	.043
16	.342	.088
19	.523	.132
22	1.372	.248
25	1.329	.195
28	1.051	.250
31	4.537	.257
34	8.886	.762
37	>8	.264
40	-	.557
43	-	1.157
46	-	.547
49	-	.890
52	-	.774
55	-	.776
58	-	1.441
61	-	.982
64	-	.867
67	-	1.492
70	-	1.361
73	-	1.322
76	-	1.363
79	-	1.647
82	-	1.109
85	-	1.407
88	-	1.691
91	-	1.623
94	-	2.043
97	-	4.736
100	-	5.007

Table 4 (continued)

m = 10 n = 20			m = 10 n = 30		
p	AL1	AL2	p	AL1	AL2
2	.131	.208	2	.307	.569
3	.456	.339	3	1.160	.718
4	1.438	.961	4	3.847	1.207
5	3.719	1.572	5	>4	4.794
6	>4	1.698	6	-	7.705
7	-	2.141	7	-	17.946
8	-	2.216	8	-	18.977
9	-	1.772	9	-	33.398
10	-	1.301			
11	-	1.119			
12	4.709	.522	22	-	5.265
13	2.174	.413	23	-	2.207
14	1.102	.345	24	21.712	.906
15	.708	.305	25	-	.731
16	.473	.303	26	-	.687
17	.318	.313	27	-	.720
18	.239	.335	28	-	.743

5. REFERENCES

- [1] Loon, J.N.M. van and J.A.M. Schreuder, ON THE MINIMAL NUMBER AND LOCATION OF FIRE STATIONS AND FIRE APPLIANCES IN ROTTERDAM, in the book: PREPRINTS/ SECOND EUROPEAN CONGRESS ON OPERATIONS RESEARCH, North-Holland, Amsterdam 1976
- [2] Tomlin, J.A., BRANCH AND BOUND METHODS FOR INTEGER AND NON-CONVEX PROGRAMMING, in the book: INTEGER AND NON-LINEAR PROGRAMMING, North-Holland, Amsterdam, 1970
- [3] Hansen, P., LES PROCEDURES D'EXPLORATION ET D'OPTIMISATION PAR SEPARATION ET EVALUATION, A SURVEY, in the book: COMBINATORIAL PROGRAMMING, METHODS AND APPLICATIONS, D. Reidel, Dordrecht, 1974
- [4] Oudheusden, D. van and F. Plastria, UN MODELE DE LOCALISATION COMME APPLICATION DE LA THEORIE DES JEUX, Working Paper CSOOTW/112, Vrije Universiteit Brussel, Brussels, 1978

USING THE TRANSPORTATION METHOD TO ALLOCATE COMBAT
AIRCRAFT SORTIES IN A HOSTILE ENVIRONMENT

BRUCE C. ELWELL, MAJOR, USAF

6168ABS/LG, 8th Tactical Fighter Wing (Taegu)
APO San Francisco 96213

ABSTRACT. The problem of determining the optimal day to day combat sortie allocations in the tactical air, air interdiction, and close air support mission roles is extremely complex and difficult. Basically the problem is one of allocating existing forces to required targets in the best possible manner (lowest total "cost"). Factors such as the overall objectives, the capabilities of each side, and the operational environment must be considered. Likewise the elements of target selection, sortie availability, desired effects, and overall costs must be entertained.

The transportation method determines a minimum cost program for "transporting" a given product or commodity from several supply locations to several demand locations. Since many routes are usually available, each with its own inherent cost, the objective is to select the most efficient, i.e., the minimum cost program.

In this paper, the allocation of combat sorties is couched in terms of the classical transportation model. The bases at which existing aircraft and ordnance (sorties) are presently stationed are viewed as the sources. The target areas at which the sorties are required, in an analogous manner, are looked upon as the destinations. The commodity to be transported is represented by the sorties.

The use of the transportation method in allocating combat sorties in a hostile environment results in a more effective and efficient use of scarce resources. In an actual combat situation this could easily be the difference between ultimate mission success or failure. Today's advanced communications technology allows for the use of such a model in a real time mode; hence, an actual application such as suggested is both feasible and practical.

1. INTRODUCTION

This paper proposes the application of an operations research technique, the transportation method, to the real world decision problem of allocating combat sorties in a hostile environment. The paper contains three major sections. The first briefly reviews the problem environment and provides insight into the allocation situation. The second section describes the transportation method, its major characteristics, and some of its classical applications. The final section integrates the first two and shows how the transportation method can be used to assist in the allocation of combat sorties in a hostile environment. A summary is also provided.

2. THE PROBLEM ENVIRONMENT

There are five basic missions of tactical air power. These are: counterair, air interdiction, close air support, tactical air reconnaissance, and tactical airlift [1:2]¹. The problem of allocating combat sorties discussed in this paper refers to the missions of air interdiction and close air support. With slight modification, however, the methodology could be applied to the other three tactical air missions as well.

The two missions of concern in this paper may be described as follows:

Air interdiction operations are conducted to destroy, neutralize, or delay the enemy's military potential before it can be brought to bear effectively against friendly forces. Targets include enemy lines of communications, bridges, personnel, vehicles, equipment and supplies beyond the immediate ground battle area and, in some cases, strikes against manufacturing, shipping, and storage areas [1:2].

1. The first number in the brackets refers to the bibliography reference number; the second number refers to the page within that reference.

Close air support attacks are conducted against hostile targets that are in close proximity to friendly forces and require detailed integration of each air mission with the fire and movement of these forces [1:2].

In essence the two missions involve the assignment of combat sorties to specific targets in order to achieve a particular goal. In the case of air interdiction, the purpose is to eliminate or hamper the foe's ability to wage war against the friendly forces. In the close air support type mission the objective is to degrade the enemy's capability by assigning aircraft to specific targets for the purpose of easing the pressure on friendly ground forces.

In a combat situation, the planning for the employment of existing forces to accomplish these two types of missions must take into account the following significant factors: (a) the specific objectives and tasks to be accomplished by the joint force commander; (b) the current characteristics of the enemy's capability; (c) the current characteristics of the friendly force's capability; and (d) the overall combat environment (weather, expectations, anticipated attrition, changes to capabilities, etc.) [1:2]. Once these basic considerations have been taken into account, the day-to-day operational strike planning effort can be made. As noted in the United States Air Force Manual 2-1, Tactical Air Operations - Counter Air, Close Air Support, and Air Interdiction, dated 2 May 1969, this is not an easy task:

Day-to-day planning for the employment of tactical air forces is a complex process of integrating capabilities and limitations in such a manner that optimum results are achieved in an ever changing tactical environment [1:12].

Under the present day concept of operations, there are two categories into which combat sortie allocations may be categorized. The first is in support of the preplanned air request which encompasses those requirements which can be foreseen and hence, permit indepth planning and coordination. The second is in support of the immediate air request [2:15]. The immediate air request, as its name

implies, does not permit such planning. This type of request is generated in a real time mode as a particular situation develops on the battle field. The allocation of combat sorties to fulfill both types of requests is accomplished by the Tactical Air Control System (TACS). Specifically the Tactical Air Control Center (TACC), which is the hub of the TACS, is primarily responsible for the satisfaction of all preplanned types of air requests [5:36]. The satisfaction of immediate air requests is the responsibility of the Direct Air Support Center (DASC), which functions as a forward element of, and is subordinate to, the TACC [7:26].

The major elements which must be considered to successfully accomplish this day-to-day employment planning by the TACC and DASC include actual target selection, weapons availability and capability, and the specific desired effect. The ease of target acquisition and identification are also important elements [1:12-13]. Of course another important consideration in this process is the "cost" of allocating each combat sortie to the required targets. All other things being equal, a sortie allocation which accomplishes the desired effects in terms of target destruction and "costs" less than a similar allocation will be preferred to that similar, but more expensive, allocation.

Hence the problem environment is one where the TACC is faced with the prospect of allocating a given number of aircraft (sorties) located at various locations to different targets within the battle area. The aircraft at these locations will undoubtedly be different in terms of types, numbers, payload capabilities, and costs of operation. Likewise the targets eligible for destruction will undoubtedly be different in terms of the number of sorties required to achieve the desired effect and relative distance from each of the friendly force operating locations. Previously mentioned was the fact that the employment planning decision process, carried out on a day-to-day basis, is not an easy task especially if one attempts to come close to an optimum solution. Even with a relatively small quantity of friendly force operating locations, number of combat sorties available for allocation from those locations, and number of targets requiring ordnance; the amount of possible allocation schemes which will satisfy the goal of meeting all target ordnance requirements is extremely large. Picking the most efficient of these schemes (in terms of overall resources consumed) is extremely difficult. As will be

shown later in this paper, the combat sortie allocation problem is amenable to solution with an operations research tool known as the "transportation method." The next section of this paper discusses the transportation method, and reviews its basic characteristics.

3. THE TRANSPORTATION METHOD

The transportation method is one of the classic techniques associated with the operations research discipline. The method has its origin in work done by F.L. Hitchcock in 1941 with a study titled "The Distribution of a Product from Several Sources to Numerous Localities" [6:213]. A parallel, but unrelated work, by T.C. Koopmans titled "Optimum Utilization of the Transportation System" was accomplished in 1947. Koopmans' work concerned the solution of a war related problem to reduce overall shipping times. The overall purpose was to eliminate a severe shortage of cargo ships constituting a critical bottleneck in the military shipping system during World War II [3:300]. These two works form the basis of the mathematical techniques which has come to be known as the transportation method [6:212]².

2. The transportation problem was first formulated in terms of a linear programming problem by George B. Dantzig in 1953 [4:311]. It was soon discovered that because of its special structure, this type of linear program was more easily solved using a modified solution procedure. There are currently many computer solution programs directed at the transportation problem which could be so utilized within the Tactical Air Control System. In an article titled "Building a Better Bubble," Major Fred Meurer describes six improvement measures designed to increase the effectiveness of the existing TACS. The most important of these, in terms of this paper, is that of totally automating the Tactical Air Control Centers. This of course would involve the computerization of many functions currently accomplished by the TACC. The point here is that with such computerization, the use of techniques such as the transportation method would become highly feasible in the TACC. For more information on this subject, see bibliography reference [5].

The transportation problem deals with the distribution of a commodity from various sources to various destinations at minimum cost. Dantzig, in his classic work titled Linear Programming and Extensions described the problem as follows:

The classical transportation problem arises when we must determine an optimal schedule of shipments that: (a) originate at sources (supply depots) where fixed stockpiles of a commodity are available; (b) are sent directly to their final destinations (demand depots) where various fixed amounts are required; (c) exhaust the stock piles and fulfill the demand, hence total demand equals total supply; and finally, the cost must (d) satisfy a linear objective function; that is, the cost of each shipment is proportional to the amount shipped, and the total cost is the sum of the individual costs [3:299].

What Dantzig is saying is this: a homogeneous product is available at various locations in fixed quantities. The availability is for a certain period of time, that is, per day, per month, etc. This commodity is required at various other locations also in fixed quantities. These requirements exist for the same period of time just noted; that is per day, per month, etc. In the final solution, the products are sent directly from the sources to the destinations; there is no transshipment possible. The total amounts of the commodity available at all sources must be equal to the total amounts of the commodity required at all destinations³. Finally, the cost of shipping each unit of the product from any source to any destination is directly

3. If in the actual problem to be solved, this is not the case, then there are ways of adding "dummy" sources or destinations to bring the problem into balance. For example, if the total amount required is less than the total amount available, then a fictitious destination requiring the differential amount is added to the problem. If the opposite is true, then a dummy source will be added which has an available quantity equal to the differential amount.

proportional to the amount shipped. For example, if it costs \$2.50 to ship one unit of the product from source A to destination B, then it will cost \$5.00 to ship two units, \$10.00 to ship four units, etc., from A to B. The optimal shipping schedule is one which accomplishes the shipments at minimum total cost.

The transportation method is used to assist in solving many different real world problems in today's commercial environment. The classical application attacks the problem of shipping a commodity from various warehouses to various retail outlets at a minimum cost. The method is also used in a production scheduling environment where the production months are the sources and the required months are the destinations. The labor cost, inventory holding costs (both for raw materials and finished goods), and miscellaneous costs are combined to form the cost of "transporting" a unit from its production month to its required month in this application. A special version of the transportation method, called the assignment method, is frequently used to allocate different jobs to different machines.

4. USING THE TRANSPORTATION METHOD TO ALLOCATE COMBAT SORTIES

The use of the transportation method as an aid to solving the combat sortie allocation problem is best demonstrated by use of an example⁴. The availability time period mentioned in chapter 3 of this paper is specified as one day and the type of mission scenario is that of satisfying preplanned requests for air interdiction. There are, for ease of presentation, three target areas (TAs) which must receive sorties; these are TAs I, II, and III respectively. It is estimated that TA I will require 72 sorties to produce the desired effect, TA II will require 102 sorties, and TA III will require 41 sorties. The friendly forces have aircraft available at three tactical unit operating locations (OLs).

4. The numbers used in the example which follows were extracted from a problem described in the Levin-Kirkpatrick text, Quantitative Approaches to Management. See bibliography reference [4], pages 328-329.

These are OLs A, B, and C. The numbers of sorties available are as follows: OL A, 76 sorties; OL B, 82 sorties; and OL C, 77 sorties. Because of the different types of aircraft involved and distances to be traveled between OLs and TAs, the unit costs per sortie between the three OLs and the three TAs have been estimated as shown in Table 1.

Table 1

UNIT COSTS PER SORTIE				
FROM	TO	TARGET AREA		
		I	II	III
Operating Location	A	4	8	8
	B	16	24	16
	C	8	16	24

Note: The entries to this table are cost units, e.g., thousands of dollars.

The objective is to allocate the available sorties to achieve the desired results at the minimum cost. One possible solution to this problem is shown in Table 2.

Table 2

Sorties from OL		to TA	Cost
72	A	I	288
4	A	II	32
82	B	II	1968
16	C	II	256
41	C	III	984
Total Cost =			3528

The optimal solution⁵ to this problem, found by using the transportation method, suggests the allocation of sorties shown in Table 3.

Table 3

Sorties from	OL	to	TA	Cost
76	A		II	608
21	B		II	504
41	B		III	656
72	C		I	576
5	C		II	80
				Total Cost = 2424

Note that the optimal solution is considerably better in terms of cost unit savings, than the initial solution. The optimal solution, however, achieves the same degree of target coverage; the only thing that is different is its cost.

Effective and efficient resource utilization is absolutely essential during any combat situation. The force which can use its existing resources to the fullest extent possible will be in a commanding position. The transportation method can be used as an aid in accomplishing that end with respect to the combat sortie allocation problem. While it only directly applies to one aspect of the combat situation, its use will help insure that combat sorties are allocated in the best manner possible thus improving the relative position of the using agency.

5. It was not the purpose or intent of this paper to discuss the solution procedures required by the transportation method. Adequate descriptions of these procedures can be found in bibliography references [3], [4], or [6].

5. SUMMARY

The transportation method is one of the classic techniques associated with the subject area of operations research. Basically, this method addresses the problem of shipping a product or commodity from various locations to various destinations at minimum total cost. In a combat environment, the allocation of aircraft sorties in support of the air interdiction and close air support mission profiles is a very complex process. This allocation process is accomplished under the purview of the Tactical Air Control System; specifically by the Tactical Air Control Center. By considering the friendly force operating locations as the sources, the enemy force target areas as the destinations and the friendly force combat sorties as the commodity to be delivered between the two, the transportation method can be used as an aid in solving the combat sortie allocation problem. The efficient use of scarce resources is especially critical in any combat situation. By using the transportation method as described in this paper, scarce resources will be more effectively and efficiently used; thus providing an important advantage to the using agency.

REFERENCES

- [1] Air Command and Staff College, Course 1F, Lesson 7, TACTICAL AIR MISSIONS, Extension Course Institute, Air University, 1976.
- [2] Air Command and Staff College, Course 1F, Lesson 10, TACTICAL AIR CONTROL SYSTEM (TACS), Extension Course Institute, Air University, 1976.
- [3] Dantzig, George B., LINEAR PROGRAMMING AND EXTENSIONS, RAND Memorandum R-366-PR. Santa Monica, California: The RAND Corporation, 1963.
- [4] Levin, Richard I. and Charles A. Kirkpatrick, QUANTITATIVE APPROACHES TO MANAGEMENT, Third Edition. New York: The McGraw Hill Book Company, 1975.
- [5] Meurer, Fred, BUILDING A BETTER BUBBLE, Air Force Magazine, April 1975, PP. 32-37.
- [6] Thierauf, Robert J. and Robert C. Klekamp, DECISION MAKING THROUGH OPERATIONS RESEARCH. Second Edition. New York: John Wiley & Sons, Inc., 1975.
- [7] Turke, C. W., TACS IN 10 MINUTES OR LESS, TAC Attack, October 1974, PP. 24-28.

OPTIMAL ALLOCATION STRATEGIES FOR
HETEROGENEOUS - FORCE DIFFERENTIAL COMBAT

HYUNG KANG SHIN and GIL CHANG KIM

The Korea Advanced Institute of Science
P.O.Box 150 Chongyangni Seoul, Korea

ABSTRACT. The paper is concerned with a method for the optimum allocation strategy of a two-on-two combat model which is an important and fundamental type of a deterministic, constant attrition-rates, Lanchester-type process between two heterogeneous forces.

A systematic procedure for the analytic solutions of the heterogeneous force differential combat equation is presented when the alternative tactics of each force are determined. The determination of optimal allocation strategies can be found by using a digital computer search technique with analytic solutions. This is accomplished by using the state increment dynamic programming developed by Larson[12] and the analytic solution for each play. The digital computer search technique by the state increment dynamic programming is an improved multistage decision process to reduce the high-speed memory requirements.

1. INTRODUCTION

Each of Blue and Red has different types of weapon systems respectively. The effectiveness of each type of weapon system on one side of combat to each type of weapon system on the other is not all the same. Such combat aspect is called "heterogeneous combat".

The two-on-two combat model in which each of Blue and Red has two different types of weapon systems is fundamental in the heterogeneous combat. A general form for the heterogeneous combat is

$$\begin{aligned}\frac{dX^B}{dt} &= A(t)X^R \\ \frac{dX^R}{dt} &= B(t)X^B\end{aligned}\tag{1}$$

where X^B and X^R are Blue and Red component vectors respectively.

Heterogeneous combat was considered initially by Helmer[6] and Snow. [13][3]

The question of optimal assignment of the weapons to the heterogeneous target of the opponent as the form of the two-on-two combat problem was mentioned by Weiss.[18] The type of the result is that the optimal tactics are either 0 or 1 as a function of the attrition rates and the force levels. He optimized the value of the game which is the difference of only the primary units of Blue and Red at the final time by using the differential game method developed by Isaacs[7].

A differential game treatment of Lanchester equation is also found in the work of Isbell and Marlow [8]. This is the case of two-on-one combat.

Taylor[14][15][16] considered fire distribution problems for a homogeneous force against heterogeneous enemy forces by using deterministic optimal control theory (Pontryagin maximum principle). The results of his works are that the optimal allocation policy is the fire concentration on one target type under a "square-law" attrition process and is sensitive to force levels, target acquisition process, the type of attrition process, and the termination conditions of combat.

Kawara[9] suggested a two-on-two combat problem under a "mixed-law" attrition process in the form of a differential game. In this model, the optimal strategy of support units is the concentration on one target type as a function of the effectiveness of both side's support units.

Weiss and Kawara considered two-on-two combat model in

which each of Blue and Red has two alternative tactics. They put great emphasis on the allocation of the support fire, thus neglecting the role of the primary units such that they can not attrite any of the enemy. But the primary unit has an important position in the tactical situation. Thus it is necessary to develop the model in which all the tactics for the primary and support units can be employed.

And the determination of optimal allocation strategies via the theory of differential game is limited to specific situations because of the computational difficulties involved. So the dynamic allocation can be accomplished by applying the dynamic programming approach.

2. FORMULATION OF HETEROGENEOUS COMBAT MODEL

The prescribed duration battle between Blue and Red, each of which has two types of weapon systems is considered. The optimization model of the constant attrition rate, heterogeneous force, differential combat can be described as follows:

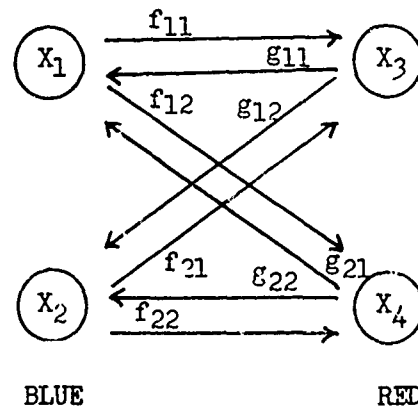


Fig. 1 The Tactical Model

$$\max_F \min_G \left\{ X^B(T) w^B \dots X^R(T) w^R \right\} \quad (2)$$

$t_0 \leq t \leq T$

subject to

$$\frac{dX}{dt} = -XA \quad (3)$$

$$A = \begin{pmatrix} 0 & C \\ D & 0 \end{pmatrix} \quad (4)$$

$$X(t=0) = (X^B(0), X^R(0)) \quad (5)$$

$$C = [C_{ij}] = [f_{ij} \beta_{ij}] \quad , i, j = 1, 2 \quad (6)$$

$$F = [f_{ij}], f_{ij} \geq 0 \quad , i, j = 1, 2 \quad (7)$$

$$\sum_j f_{ij} = 1 \quad , i = 1, 2 \quad (8)$$

$$D = [d_{ij}] = [g_{ij} \gamma_{ij}] \quad , i, j = 1, 2 \quad (9)$$

$$G = [g_{ij}], g_{ij} \geq 0 \quad , i, j = 1, 2 \quad (10)$$

$$\sum_j g_{ij} = 1 \quad , i = 1, 2 \quad (11)$$

$$X_i^- \leq X_i \leq X_i(0) \quad , i = 1, 2, 3, 4 \quad (12)$$

$$T \leq t_f \quad (13)$$

In the above model all symbols are defined as follows :

$X = (X^B, X^R)$ is the component vector

where $X^B = (X_1, X_2)$ is the Blue component and $X^R = (X_3, X_4)$

is the Red component, $X^B(0)$ and $X^R(0)$ are the initial values

β_{ij} is the attrition-rate coefficient - the rate at which each weapon system in the i -th Blue component attrites the j -th Red component.

f_{ij} is the allocation factor - the fraction of the i -th Blue component assigned to the j -th Red component. f_{ij} is either 0 or 1.

γ_{ij}, g_{ij} can be defined similarly.

$W = (W^B, W^R)^T$ is the weight vector.

where $W^B = (W_1, W_2)^T$ is the Blue weight vector and $W^R = (W_3, W_4)^T$ is the Red weight vector.

t_f is the final time of combat. X_i^- is the lower limit of X_i called the defeat criterion.

Each weapon type of Blue allocates all of his power to only one Red weapon type at any fixed time, vice versa. And during the combat each weapon type loses all of his capability at the defeat criterion, then this weapon type is deleted from the combat. The combat is terminated when $t = t_f$ or when either side loses all of his capabilities. T is the time satisfying one of these final conditions.

3. PROCEDURE FOR THE ANALYTIC SOLUTIONS OF THE DIFFERENTIAL COMBAT EQUATION

In this section a systematic procedure for the analytic solution of the eqs.(3) and (5) is presented.

Let $A(t)$ be a continuous function from the interval (T_1, T_2) into the set of $n \times n$ matrices.

Suppose that $X(0) \in R_n, t_0 \in (T_1, T_2)$.

Then there exists a function $\psi(t)$ from all of (T_1, T_2) into R_n such that

- (1) $\psi(t)$ is continuous.
- (2) $\psi(t_0) = X(0)$.
- (3) $\psi(t)$ is a solution of the linear system

$$\frac{dX}{dt} = A(t) X(t).$$
- (4) $\psi(t)$ is unique.

And the unique solution is

$$X(t) = X(0) e^{A(t-t_0)}. \quad [1] \quad (14)$$

If $t_0 = 0$,

$$X(t) = X(0) e^{At} \quad (15)$$

$$= X(0) \sum_{k=0}^{\infty} A^k \cdot \frac{t^k}{k!} \quad (16)$$

The matrix A is similar to a matrix of Jordan canonical form J [10] and e^{At} and e^{Jt} are similar matrices [1], i.e.,

$$e^{Jt} = P e^{At} P^{-1} \quad \text{if } J = P A P^{-1} \quad (17)$$

If for every characteristic value λ of A , $|\lambda| < r$ where r is the radius of convergence of $\sum_{k=0}^{\infty} a_k X^k$, then the series $\sum_{k=0}^{\infty} a_k (\lambda I + N)^k$ converges where N is nilpotent matrix. [5]

Thus the analytic solution of the heterogeneous-force differential equation can be expressed as

$$X(t) = X(0) P^{-1} e^{Jt} P \quad (18)$$

A systematic procedure is as follows:

- (1) Determine the distinct characteristic values $\lambda_1, \dots, \lambda_j$ and their multiplicity s_1, \dots, s_j of the matrix A .
- (2) Calculate the number c_i of linearly independent characteristic vectors that corresponds to λ_i .
- (3) If $c_i = s_i$, then $J_i = \lambda_i I$. And the characteristic vectors can be easily obtained.
- (4) If $c_i < s_i$, the $s_i - c_i$ is the number of 1's on the superdiagonal of J_i .
- (5) Determine the Jordan canonical matrix J where

$$J = \begin{bmatrix} J_1 & & 0 \\ & J_i & \\ 0 & & J_j \end{bmatrix} \quad \text{and} \quad J_i = \begin{bmatrix} \lambda_i & 1 & & \\ & \lambda_i & \ddots & \\ & & \ddots & 1 \\ & & & \lambda_i \end{bmatrix} \quad (19)$$

- (6) Determine the matrix P_i by the matrix equations

$$\begin{aligned}
P_{i,m} (A - \lambda_i I) &= 0 \\
P_{i,m-1} (A - \lambda_i I)^2 &= 0 \\
&\vdots \\
P_{i1} (A - \lambda_i I) &= 0
\end{aligned}
\tag{20}$$

Then $P = \begin{bmatrix} P_1 \\ \vdots \\ P_j \end{bmatrix}$ where $P_i = \begin{bmatrix} P_{i1} \\ \vdots \\ P_{im} \end{bmatrix}$, P_{ij} is a row vector.

(7) Calculate the submatrices

$$S(J_i) = \begin{bmatrix} S(\lambda_i) & S^{(1)}(\lambda_i) & S^{(2)}(\lambda_i)/2! & \dots & S^{(P-1)}(\lambda_i)/(P-1)! \\ & S(\lambda_i) & S^{(1)}(\lambda_i) & \dots & S^{(P-2)}(\lambda_i)/(P-2)! \\ & & \ddots & \ddots & \vdots \\ 0 & & & \ddots & S(\lambda_i) \end{bmatrix} \tag{21}$$

$$\text{where } S(\lambda_i) = \sum_{k=0}^{\infty} (-1)^k (\lambda_i)^k t^k/k!$$

and $S^{(r)}(\lambda_i)$ is the r -th derivative of $S(\lambda_i)$.

$$\tag{22}$$

(8) Then the analytic solution is

$$X(t) = X(0) P^{-1} S(J) P$$

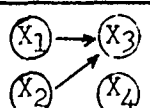
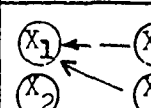
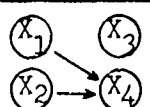
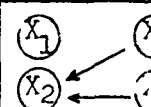
where

$$S(J) = \begin{bmatrix} S(J_1) & \dots & 0 \\ 0 & \dots & S(J_j) \end{bmatrix} \tag{23}$$

.. ANALYSIS OF THE TWO-ON-TWO COMBAT

The alternative tactics of Blue and Red in the two-on-two combat can be represented by the Table 1.

Table 1. Alternative tactics of Blue and Red.

Blue		Red	
alterna- tive	tactic	alterna- tive	tactic
1		1	
2		2	

Blue		Red	
alternat- ive	tactic	alternat- ive	tactic
3	$\begin{array}{c} \textcircled{X_1} \rightarrow \textcircled{X_3} \\ \textcircled{X_2} \rightarrow \textcircled{X_4} \end{array}$	3	$\begin{array}{c} \textcircled{X_1} \leftarrow \textcircled{X_3} \\ \textcircled{X_2} \leftarrow \textcircled{X_4} \end{array}$
4	$\begin{array}{c} \textcircled{X_1} \rightarrow \textcircled{X_3} \\ \textcircled{X_2} \rightarrow \textcircled{X_4} \end{array}$	4	$\begin{array}{c} \textcircled{X_1} \leftarrow \textcircled{X_3} \\ \textcircled{X_2} \leftarrow \textcircled{X_4} \end{array}$

The two-on-two combat is developed by the simultaneous choices of tactics for Blue and Red. An ordered pair (i, j) where i is one of Blue's alternatives and j is one of Red's alternatives, is called a play and denoted as $p(i, j)$. The set of all the plays can be denoted as

$$P = \{p(i, j) ; i, j \in \{1, 2, 3, 4\}\} \quad (24)$$

Then a two person, zero-sum game is organized [17]: (1) There are two players designated Blue and Red. (2) Each of Blue and Red has 4 alternative tactics, (3) A play of game occurs when each player chooses an alternative simultaneously. (5) As a result of the play $p(i, j)$, there is a payment e_{ij} from Red to Blue. The computation of the payment will be explained in the next section. Let $B = (b_1, b_2, b_3, b_4)$ correspond to the probability that Blue will use alternatives (1, 2, 3, 4) respectively and $R = (r_1, r_2, r_3, r_4)$ correspond to the probability that Red will use alternatives (1, 2, 3, 4) respectively. The vectors B and R are called strategies for Blue and Red and these are generally mixed strategies. Given the strategy B and R , the expected value from a play of the game is

$$E(B, R) = \sum_i \sum_j e_{ij} b_i r_j. \quad (25)$$

Blue's objective is to maximize $E(B, R)$ and Red's objective is to minimize $E(B, R)$. Then there exists an optimal solution (B^*, R^*) such that

$$\max_B \min_R E(B, R) = \min_R \max_B E(B, R) = E(B^*, R^*) [4] \quad (26)$$

The solution of this game can be obtained easily by linear programming. [4]

The set of all the plays for the two-on-two combat can be partitioned into 7 types. For every type, all the plays and the analytic solution of the first play in every type is presented. The solutions of the other plays in the same type can be obtained by permuting the indices of the first play. For example, the solution of the play $P(2, 2)$ of type I can be

obtained by the permutation $\begin{pmatrix} 1 & 2 & 3 & 4 \\ 2 & 1 & 4 & 3 \end{pmatrix}$ of all the indices in eq.(27).

Table 2. Plays and solution of Type I

Type	Type I			
	P(1,1)	P(1,2)	P(2,1)	P(2,2)
Play				
Permutation	$\begin{pmatrix} 1 & 2 & 3 & 4 \\ 1 & 2 & 3 & 4 \end{pmatrix}$	$\begin{pmatrix} 1 & 2 & 3 & 4 \\ 2 & 1 & 3 & 4 \end{pmatrix}$	$\begin{pmatrix} 1 & 2 & 3 & 4 \\ 1 & 2 & 4 & 3 \end{pmatrix}$	$\begin{pmatrix} 1 & 2 & 3 & 4 \\ 2 & 1 & 4 & 3 \end{pmatrix}$
Y_{11}	$\begin{bmatrix} X_1(t) & X_2(t) & X_3(t) & X_4(t) \end{bmatrix} = \begin{bmatrix} X_1(0) & X_2(0) & X_3(0) & X_4(0) \end{bmatrix} \begin{bmatrix} \cosh \sqrt{a_{13}a_{31}}t & 0 & -\frac{a_{13}}{\sqrt{a_{31}}} \sinh \sqrt{a_{13}a_{31}}t & 0 \\ \frac{a_{23}}{a_{13}} (\cosh \sqrt{a_{13}a_{31}}t - 1) & 1 - \frac{a_{23}}{\sqrt{a_{13}a_{31}}} \sinh \sqrt{a_{13}a_{31}}t & 0 & 0 \\ -\frac{\sqrt{a_{31}}}{a_{13}} \sinh \sqrt{a_{13}a_{31}}t & 0 & \cosh \sqrt{a_{13}a_{31}}t & 0 \\ -\frac{a_{41}}{\sqrt{a_{13}a_{31}}} \sinh \sqrt{a_{13}a_{31}}t & 0 & \frac{a_{41}}{a_{31}} (\cosh \sqrt{a_{13}a_{31}}t - 1) & 1 \end{bmatrix} \quad (27)$			

Table 3. Plays and solution of Type II

Type	Type II			
	P (1,3)	P (1,4)	P (2,3)	P (2,4)
Play				

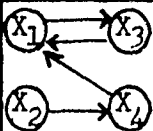
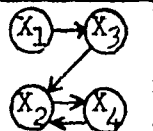
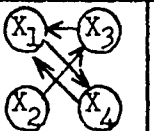
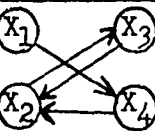
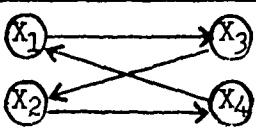
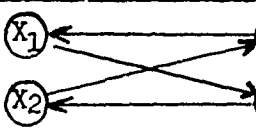
Permutation	$\begin{pmatrix} 1 & 2 & 3 & 4 \\ 1 & 2 & 3 & 4 \end{pmatrix}$	$\begin{pmatrix} 1 & 2 & 3 & 4 \\ 2 & 1 & 3 & 4 \end{pmatrix}$	$\begin{pmatrix} 1 & 2 & 3 & 4 \\ 2 & 1 & 4 & 3 \end{pmatrix}$	$\begin{pmatrix} 1 & 2 & 3 & 4 \\ 1 & 2 & 4 & 3 \end{pmatrix}$
	P (3,1)	P (3,2)	P (4,1)	P (4,2)
Play				
Permutation	$\begin{pmatrix} 1 & 2 & 3 & 4 \\ 3 & 4 & 1 & 2 \end{pmatrix}$	$\begin{pmatrix} 1 & 2 & 3 & 4 \\ 4 & 3 & 2 & 1 \end{pmatrix}$	$\begin{pmatrix} 1 & 2 & 3 & 4 \\ 4 & 3 & 1 & 2 \end{pmatrix}$	$\begin{pmatrix} 1 & 2 & 3 & 4 \\ 3 & 4 & 2 & 1 \end{pmatrix}$
Y_{13}	$ \begin{aligned} & [X_1(t) \ X_2(t) \ X_3(t) \ X_4(t)] = [X_1(0) \ X_2(0) \ X_3(0) \ X_4(0)] \\ & \begin{pmatrix} \cosh \sqrt{a_{13}a_{31}} t & 0 & -\frac{a_{13}}{\sqrt{a_{31}}} \sinh \sqrt{a_{13}a_{31}} t & 0 \\ \frac{a_{23}}{a_{13}} (\cosh \sqrt{a_{13}a_{31}} t - 1) & 1 & \frac{-a_{23}}{\sqrt{a_{13}a_{31}}} \sinh \sqrt{a_{13}a_{31}} t & 0 \\ -\frac{\sqrt{a_{31}}}{\sqrt{a_{13}}} \sinh \sqrt{a_{13}a_{31}} t & 0 & \cosh \sqrt{a_{13}a_{31}} t & 0 \\ \frac{a_{23}a_{42}}{a_{13}} (t - \sinh \frac{\sqrt{a_{13}a_{31}} t}{\sqrt{a_{13}a_{31}}}) & -a_{42} t & \frac{a_{23}a_{42}}{a_{13}a_{31}} (\cosh \sqrt{a_{13}a_{31}} t - 1) & 1 \end{pmatrix} \quad (28) \end{aligned} $			

Table 4. Plays and solution of Type III

Type	Type III	
	P (3,4)	P (4,3)
Play		
Permutation	$\begin{pmatrix} 1 & 2 & 3 & 4 \\ 1 & 2 & 3 & 4 \end{pmatrix}$	$\begin{pmatrix} 1 & 2 & 3 & 4 \\ 1 & 2 & 4 & 3 \end{pmatrix}$

$$\begin{aligned}
 Y_{34} [X_1(t) \ X_2(t) \ X_3(t) \ X_4(t)] &= [X_1(0) \ X_2(0) \ X_3(0) \ X_4(0)] \\
 &\begin{aligned}
 &\frac{1}{2}(\cosh k_2 t + \cosh k_2 t) \frac{a_{13}a_{32}}{2k_1} \cosh k_2 t - \frac{a_{13}}{2k_2} \sinh k_2 t - \frac{a_{13}a_{32}a_{24}}{2k_1k_2} (\sinh k_2 t - \sinh k_2 t) \\
 &\frac{a_{24}a_{41}}{2k_1} \cosh k_2 t - \frac{1}{2}(\cosh k_2 t + \cosh k_2 t) \frac{a_{24}a_{41}a_{13}}{2k_1k_2} \sinh k_2 t - \frac{a_{24}}{2k_2} (\sinh k_2 t - \sinh k_2 t) \\
 &-\frac{a_{32}a_{24}a_{41}}{2k_1k_2} (\sinh k_2 t - \sinh k_2 t) - \frac{a_{32}}{2k_2} (\sinh k_2 t + \sinh k_2 t) + \frac{1}{2}(\cosh k_2 t + \cosh k_2 t) \frac{a_{32}a_{24}}{2k_1} (\cosh k_2 t - \cosh k_2 t) \\
 &-\frac{a_{41}}{2k_2} (\sinh k_2 t + \sinh k_2 t) - \frac{a_{13}a_{32}a_{41}}{2k_1k_2} (\sinh k_2 t - \sinh k_2 t) + \frac{a_{13}a_{41}}{2k_1} (\cosh k_2 t - \cosh k_2 t) + \frac{1}{2}(\cosh k_2 t + \cosh k_2 t) \\
 &\frac{1}{2k_2} (\sinh k_2 t + \sinh k_2 t)
 \end{aligned} \\
 \text{where } K &= a_{13}a_{32}a_{24}a_{41} \\
 K_1 &= K^{\frac{1}{2}}, K_2 = K^{\frac{1}{4}} \quad (29)
 \end{aligned}$$

Table 5. Plays and solution of type IV

Type	Type IV	
	P (3,3)	P (4,4)
	$ \begin{array}{ccc} (X_1) & \longleftrightarrow & (X_3) \\ (X_2) & \longleftrightarrow & (X_4) \end{array} $	$ \begin{array}{ccc} (X_1) & \begin{array}{c} \nearrow \searrow \\ \nwarrow \nearrow \end{array} & (X_3) \\ (X_2) & \begin{array}{c} \nearrow \searrow \\ \nwarrow \nearrow \end{array} & (X_4) \end{array} $
Permu- tation	$\begin{pmatrix} 1 & 2 & 3 & 4 \\ 1 & 2 & 3 & 4 \end{pmatrix}$	$\begin{pmatrix} 1 & 2 & 3 & 4 \\ 1 & 2 & 4 & 3 \end{pmatrix}$
Y_{33}	$ \begin{aligned} [X_1(t) \ X_2(t) \ X_3(t) \ X_4(t)] &= [X_1(0) \ X_2(0) \ X_3(0) \ X_4(0)] \\ &\begin{pmatrix} \cosh \sqrt{a_{13}a_{31}}t & 0 & -\frac{\sqrt{a_{13}}}{\sqrt{a_{31}}} \sinh \sqrt{a_{13}a_{31}}t & 0 \\ 0 & \cosh \sqrt{a_{42}a_{24}}t & 0 & -\frac{\sqrt{a_{42}}}{\sqrt{a_{24}}} \sinh \sqrt{a_{42}a_{24}}t \\ -\frac{\sqrt{a_{31}}}{\sqrt{a_{13}}} \sinh \sqrt{a_{13}a_{31}}t & 0 & \cosh \sqrt{a_{13}a_{31}}t & 0 \\ 0 & -\frac{\sqrt{a_{42}}}{\sqrt{a_{24}}} \sinh \sqrt{a_{42}a_{24}}t & 0 & \cosh \sqrt{a_{42}a_{24}}t \end{pmatrix} \quad (30) \end{aligned} $	

Table 6. Plays and solution of type V

Type	Type V			
	P(1,1)	P(1,2)	P(2,1)	P(2,2)
Play				
Permutation	$\begin{pmatrix} 1 & 2 & 3 & 4 \\ 1 & 2 & 3 & 4 \end{pmatrix}$	$\begin{pmatrix} 1 & 2 & 3 & 4 \\ 2 & 1 & 3 & 4 \end{pmatrix}$	$\begin{pmatrix} 1 & 2 & 3 & 4 \\ 1 & 2 & 4 & 3 \end{pmatrix}$	$\begin{pmatrix} 1 & 2 & 3 & 4 \\ 2 & 1 & 4 & 3 \end{pmatrix}$
Y_{11}	$\begin{bmatrix} X_1(t) & X_2(t) & X_3(t) & X_4(t) \end{bmatrix} = \begin{bmatrix} X_1(0) & X_2(0) & X_3(0) & X_4(0) \end{bmatrix} \begin{pmatrix} \cosh \sqrt{a_{13}a_{31}}t & 0 & -\frac{a_{13}}{a_{31}} \sinh \sqrt{a_{13}a_{31}}t & 0 \\ \frac{a_{23}}{a_{13}} (\cosh \sqrt{a_{13}a_{31}}t - 1) & -\frac{a_{23}}{\sqrt{a_{13}a_{31}}} \sinh \sqrt{a_{13}a_{31}}t & 0 & 0 \\ -\frac{a_{31}}{a_{13}} \sinh \sqrt{a_{13}a_{31}}t & 0 & \cosh \sqrt{a_{13}a_{31}}t & 0 \\ 0 & 0 & 0 & 0 \end{pmatrix} \quad (31)$			

Table 7. Plays and solution of Type VI

Type	Type VI			
	P(1,1)	P(1,2)	P(2,1)	P(2,2)
Play				
Permutation	$\begin{pmatrix} 1 & 2 & 3 & 4 \\ 1 & 2 & 3 & 4 \end{pmatrix}$	$\begin{pmatrix} 1 & 2 & 3 & 4 \\ 2 & 1 & 3 & 4 \end{pmatrix}$	$\begin{pmatrix} 1 & 2 & 3 & 4 \\ 1 & 2 & 4 & 3 \end{pmatrix}$	$\begin{pmatrix} 1 & 2 & 3 & 4 \\ 2 & 1 & 4 & 3 \end{pmatrix}$

$$\begin{array}{c} Y_{11} \\ \left[\begin{array}{cccc} X_1(t) & X_2(t) & X_3(t) & X_4(t) \end{array} \right] = \left[\begin{array}{cccc} X_1(0) & X_2(0) & X_3(0) & X_4(0) \end{array} \right] \\ \left[\begin{array}{cccc} \cosh \sqrt{a_{31}a_{13}}t & 0 & -\frac{\sqrt{a_{13}}}{\sqrt{a_{31}}} \sinh \sqrt{a_{31}a_{13}}t & 0 \\ 0 & 1 & 0 & 0 \\ -\frac{\sqrt{a_{31}}}{\sqrt{a_{13}}} \sinh \sqrt{a_{31}a_{13}}t & 0 & \cosh \sqrt{a_{31}a_{13}}t & 0 \\ \frac{-a_{41}}{\sqrt{a_{31}a_{13}}} \sinh \sqrt{a_{31}a_{13}}t & 0 & \frac{a_{41}}{a_{31}} (\cosh \sqrt{a_{31}a_{13}}t - 1) & 1 \end{array} \right] \end{array} \quad (32)$$

Table 8. Plays and solution of type VII

Type	Type VII			
	P(1,1)	P(1,2)	P(2,1)	P(2,2)
Play				
Permutation	$\begin{pmatrix} 1 & 2 & 3 & 4 \\ 1 & 2 & 3 & 4 \end{pmatrix}$	$\begin{pmatrix} 1 & 2 & 3 & 4 \\ 2 & 1 & 3 & 4 \end{pmatrix}$	$\begin{pmatrix} 1 & 2 & 3 & 4 \\ 1 & 2 & 4 & 3 \end{pmatrix}$	$\begin{pmatrix} 1 & 2 & 3 & 4 \\ 2 & 1 & 4 & 3 \end{pmatrix}$
Y_{11}	$ \left[\begin{array}{cccc} X_1(t) & X_2(t) & X_3(t) & X_4(t) \end{array} \right] = \left[\begin{array}{cccc} X_1(0) & X_2(0) & X_3(0) & X_4(0) \end{array} \right] \\ \left[\begin{array}{cccc} \cosh \sqrt{a_{13}a_{31}}t & 0 & \frac{\sqrt{a_{13}}}{\sqrt{a_{31}}} \sinh \sqrt{a_{13}a_{31}}t & 0 \\ 0 & 1 & 0 & 0 \\ -\frac{\sqrt{a_{31}}}{\sqrt{a_{13}}} \sinh \sqrt{a_{13}a_{31}}t & 0 & \cosh \sqrt{a_{13}a_{31}}t & 0 \\ 0 & 0 & 0 & 1 \end{array} \right] \end{array} \quad (33) $			

4. DYNAMIC PROGRAMMING FORMULATION OF HETEROGENEOUS COMBAT MODEL

In the section 2, the optimization model of the constant attrition-rate, heterogeneous-force, differential combat has been suggested. In this section the model is formulated suitably for the multistage decision system which the optimal solution is established one stage at a time. In order to apply the computational procedure of this system, the variables are to be quantized.

In the original model, the eq.(2) can be expressed as the following without loss of its initial meaning

$$\max_F \min_G \left\{ (x_R(0) - x_R(T))w_R - (x_B(0) - x_B(T))w_B \right\} \quad (34)$$

$$t_0 \leq t \leq T$$

The stage variable t is continuous, but in order to implement on digital computer, it is quantized into increment denoted as Δt . The quantized value of t can be indexed by the discrete sequence,

$$k = 0, 1, \dots, K-1 \quad (35)$$

$$\text{where } t_k = t_0 + k \Delta t \text{ and } K \Delta t = t_f - t_0$$

The probability that Blue will use the alternative i is b_i and the probability that Red will use the alternative j is r_j . Then Blue's strategy is $B = (b_1, b_2, b_3, b_4)$ and Red's strategy is $R = (r_1, r_2, r_3, r_4)$ and $U = (B, R) = (b_1, b_2, b_3, b_4, r_1, r_2, r_3, r_4)$ (36) is the control vector. The control vector for play $p(i, j)$ is that b_i and r_j are 1's and others are zero's. Thus every play corresponds to the quantized point of control variable.

In the section 4, the system equations for the model are obtained. These present how the state variables of stage $k + 1$ are related to the state variables of stage k and the control variables at stage k . These equations can be expressed as

$$X(i, j)(t_k + \Delta t) = Y_{ij}(X(t_k), t_k, \Delta t) \quad (37)$$

where $X(i, j)(t_k + \Delta t)$ is the state vector at the time $t_k + \Delta t$ when the play $p(i, j)$ is continued for the time interval Δt . The probability that play $p(i, j)$ will be chosen is $b_i r_j$, thus

$$X(t_k + \Delta t) = \sum_i \sum_j Y_{ij}(X(t_k), t_k, \Delta t) b_i r_j \quad (38)$$

$$= B Y(X(t_k), t_k, \Delta t) R^T \quad (39)$$

$$= Q[X(t_k), U(t_k), t_k] \quad (40)$$

$$\text{, i.e., } X(k+1) = Q[X(k), U(k), k] \quad (41)$$

where the matrix $Y = \{Y_{ij}\}$, $i, j = 1, 2, 3, 4$

The performance criterion provides an evaluation of a given control sequence $U(0), U(1), \dots, U(K-1)$. For Blue, it is to be maximized, and for Red it is to be minimized. It depends upon each value of $U(k)$, $k=0, \dots, K-1$ and also upon each value of the state vector $X(k)$, $k=0, \dots, K-1$. Let e_{ij} be the payment from Red to Blue as a result of the play $p(i, j)$ for the time interval Δt , then

$$e_{ij} = [w^B, w^R] \begin{bmatrix} X^B(i, j)(t_k + \Delta t) - X^B(t_k) \\ X^R(t_k) - X^R(i, j)(t_k + \Delta t) \end{bmatrix} \quad (42)$$

If the criterion is denoted as S , it can be written as

$$S = \sum_{k=1}^{K-1} h[X(k), U(k), k] \quad (43)$$

where

$$h[X(k), U(k), k] = \sum_i \sum_j e_{ij} b_i r_j \quad (44)$$

Then the eq. (34) can be expressed as

$$\max_{B(k)} \min_{R(k)} \sum_{k=0}^{K-1} h[X(k), U(k), k] \quad (45)$$

The constraints place restrictions on the values that the state variable and the control variables can assume. The state vector at the stage k is constrained to be in the set

$$X(k) = \{(X_1(k), X_2(k), X_3(k), X_4(k)) : \\ X_i^- \leq X_i(k) \leq X_i(0), i=1, 2, 3, 4\} \quad (46)$$

The control vector at state X , stage k is constrained to be in the set

$$U(X, k) = \{(b_1, b_2, b_3, b_4, r_1, r_2, r_3, r_4) : \\ \sum_i b_i = 1, \quad b_i \geq 0 \\ \sum_j r_j = 1, \quad r_j \geq 0\} \quad (47)$$

Then the optimization model can be stated as follows:

Given

(i) A system described by equation (40)

(ii) Constraints described by equations (46) and (47)
 (iii) An initial condition $X(0)$
 Find :
 the control sequence $U(0), \dots, U(K-1)$
 that optimizes in equation (43)
 while satisfying the constraints.

Then the iterative functional equation for determining optimal control is obtained. At any stage k , $0 \leq k \leq K-1$, let

$$I[X(k), k] = \max_{j=k, \dots, K-1} \min_{B(j)} \sum_{R(j)}^{K-1} h[X(j), U(j), j] \quad (48)$$

The summation inside the braces can be split into two parts.

$$I[X(k), k] = \max_{u(k) \in U} \min_{u(j) \in U} \left\{ h[X(k), U(k), k] + \sum_{j=k+1}^{K-1} h[X(j), U(j), j] \right\}$$

The first term in the braces in eq. (49) depends only on $U(k)$, and the second term does not depend explicitly on $U(k)$.

$$I[X(k), k] = \max_{u(k) \in U} \min \left\{ h[X(k), U(k), k] + \max_{\substack{U(j) \in U \\ j=k+1, \dots, K-1}} \min \sum_{j=k+1}^{K-1} h[X(j), U(j), j] \right\}^{(50)}$$

The second term in eq. (50) can be expressed as eq. (51) by using eq. (41) and eq. (48)

$$\max_{\substack{U(j) \in U \\ j=k+1, \dots, K-1}} \min \sum_{j=k+1}^{K-1} h[X(j), U(j), j] = I[Q[X(k), U(k), k], k+1] \quad (51)$$

Then the functional equation can be written as

$$I[X(k), k] = \max_{u(k) \in U} \min \left\{ h[X(k), U(k), k] + I[Q[X(k), U(k), k], k+1] \right\} \quad (52)$$

This iterative functional equation describes the principle of optimality [2], i.e., the optimum value at state X and stage k can be obtained by optimizing the sum of the value from the resulting state at the next stage, $k+1$, to the end of this process. Thus the iterative functional equation for dynamic programming is

$$I[X(K-1), K-1] = \max_{U(K-1)} \min h[X(K-1), U(K-1), K] \quad (53)$$

$$S[X(k), U(k), k] = h[X(k), U(k), k] + I[X(k+1), k+1] \quad (54)$$

$k = 0, 1, \dots, K-2$

subject to

$$X(k+1) = Q[X(k), U(k), k] \quad (55)$$

$$I[X(k), k] = \max_{U(k)} \min S[X(k), U(k), k] \quad (56)$$

The basic scheme of dynamic programming is to find the sequence

$$I[X(K-1), K-1], \dots, I[X(1), 1], I[X(0), 0]$$

using equations (53) to (56). Consequently, the optimal solution for the K-stage system is established on stage at a time.

5. QUANTIZATION, BLOCK AND COMPUTATIONAL PROCEDURE

Within the range determined by eq. (12), each state variable is quantized in uniform increment ΔX_i .

$$\begin{aligned} X_{ij} &= X_i^- + j \Delta X_i \\ j &= 0, 1, \dots, N_i \\ N_i \Delta X_i &= X_i(0) - X_i^- \quad i = 1, 2, 3, 4 \end{aligned} \quad (57)$$

To obtain the payoff matrix for the combat game, there must be a finite number of play $P(i, j)$'s

$$P = \{p(i, j) : i, j = 1, 2, 3, 4\}.$$

A fundamental difference between state increment dynamic programming [10][11] and conventional dynamic programming is in the method for determining δt , the time interval over which a given control is applied. Within the conventional method this interval is a fixed value Δt . In the state increment dynamic programming, for any play the interval δt is determined as the minimum time interval required for any one of the state variables to change by one increment or Δt , i.e.,

$$\delta t = \min_{i=1,2,3,4} \left\{ \left| \frac{\Delta X_i}{dX_i/dt} \right|, \Delta t \right\} \quad (58)$$

where $\frac{dX_i}{dt}$ is the i -th component of $\frac{dX}{dt}$

in the system differential equation, eq. (3).

Then the next state lies within a small neighborhood of a present state. Thus in order to perform the interpolation of the optimal value, it is necessary to store optimal values at only these quantized states near the present state.

The reduction of the high speed memory requirements is ac-

nieved mainly by the partitioning of t - X space into rectangular units called "blocks". Each block covers W_0 increments along the t -axis and W_i increments along the X_i -axis. The block $B(j_0, j_1, j_2, j_3, j_4)$ contains a set of points (t, X) ,

$$B(j_0, j_1, j_2, j_3, j_4) = \{ (t, X_1, X_2, X_3, X_4) : \begin{aligned} &(j_0 - 1) W_0 \Delta t \leq t - t_0 \leq j_0 W_0 \Delta t, \\ &(j_i - 1) W_i \Delta X_i \leq X_i - X_i^- \leq j_i W_i \Delta X_i, \quad i=1, 2, 3, 4 \\ &\text{where } j_0 = 1, 2, \dots, J_0, \\ &J_0 W_0 \Delta t = t_f - t_0, \\ &j_i = 1, 2, \dots, J_i, \\ &J_i W_i \Delta X_i = X_i(0) - X_i^-, \quad i=1, 2, 3, 4 \end{aligned} \} \quad (59)$$

As indicated in eq.(58), the boundaries between blocks are considered to be in both blocks. The numbers W_i are taken to be small integers according to the capacity of the high speed memory. The computation is performed block by block according to the lexicographic order. By the definition of δt , the next state for any play must always lie in the same block as the state at which the control is applied.

The computation within one block is performed according to the lexicographical order. The operations at a given quantized point (t, X) take place as follows. Let $\delta t(i, j)$ be the time interval computed by eq.(58) for play $p(i, j)$. For every play $p(i, j) \in P$, $\delta t(i, j)$ is determined. The corresponding next state is computed by eqs. (27) to (33) in the section 4. The next state as a result of to play $p(i, j)$ can be stated as following expression.

$$X(i, j) (t + \delta t) = X(t) + \delta X(i, j) \quad (60)$$

The optimal value at the next state for every play is computed by interpolation in 3-dimensional state variable and time using the previously computed values at quantized states at time $t + \Delta t$. The state variable not used in the interpolation is the one for which $\delta t(i, j)$ in eq. (58). takes on the minimum value. If $\delta t(i, j) = \Delta t$, the optimal value at this point is computed by interpolation in 4-dimensional state variable. Once the optimal value $I[X(i, j), t + \delta t(i, j)]$ for the next state has been computed, the return of this control over the time interval $\delta t(i, j)$ is computed by the eq. (42), i.e.,

$$e_{ij} = [W^B, W^R] \begin{bmatrix} -\delta X^B(i, j) \\ \delta X^R(i, j) \end{bmatrix}$$

Then the optimal value for the point (t, X) as a result of the play $p(i, j)$, say $S(i, j)$

$$S(i,j)[X(t), U_{ij}(t), t] = e_{ij} + I[X_{ij}, t + \delta t(i,j)] \quad (61)$$

where U_{ij} is the control vector that b_i, r_j are 1's and others zeros.

$S(i,j)$ is the (i,j) -th element of the payoff matrix at this quantized point (t,X) . Then for a given point (t,X) a payoff matrix for a two-person, zero-sum game can be obtained by using the above method for every play $p(i,j)$, $i,j = 1,2,3,4$.

$$\text{Then } I[X(k),k] = \max_i \min_j S(i,j)[X(k), U(k),k]$$

is the value of the game, i.e.,

$$I[X(k),k] = \max_B \min_R \sum_i \sum_j S(i,j) b_i r_j \quad (62)$$

and the corresponding optimal control is $U^*(k) = (B^*, R^*)$

The optimal control and the optimal value for every quantized point in this given block has been computed in lexicographic order and then all the results of this block are stored in a low-speed memory.

When all of computations has been performed for blocks in $t_f - W_0 \Delta t \leq t \leq t_f$, the set of blocks in $t_f - 2W_0 \Delta t \leq t \leq t_f - W_0 \Delta t$ is processed. The same procedure are used, except that the optimal values just computed at $(t_f - W_0 \Delta t)$ are used to initialize these blocks. These computations continue until $t = t_0$ is reached.

The original problem is to find the optimum sequence of controls starting from the given $X(0)$. Let $U^*(X(0), 0)$ be the optimal control for the point (t,X) where $t=0$ and $X=X(0)$. The first optimal control in the sequence is evaluated as

$$U^*(0) = U^*(X(0), 0) \quad (63)$$

and the time interval over which the first control is applied is determined as the minimum time interval required for any one of the state variables and time to change by next quantized value, i.e.,

$$\delta t(U^*(i=0)) = \min_{k=1,2,3,4} \left\{ \frac{X_k(i=0) - X_k(N_i)}{[dX_k/dt]_{X(0)}}, t(N_i) - t(i=0) \right\} \quad (64)$$

where $X_k(N_i)$ is the next quantized value of X_k smaller than $X_k(i)$

and $t(N_i)$ is the next quantized value larger than $t(i)$ and $X_k(i)$ means

the initial state to which the i -th control is applied and $t(i)$ means the initial time to which the i -th control is applied.

Then the next state along the sequence

$$X^*(1) = Q[X(0), U^*(0), t(0), t(U^*(0))] \quad (65)$$

is computed by eq. (39).

The superscript * means "optimal", for example $t(U^*(0))$ mean the time interval when the optimal control $U^*(0)$ is applied. Thus the optimal value over time interval $\delta t(U^*(0))$ is

$$S^*(0) = [w^B, w^R] \begin{bmatrix} -\delta X^{*B}(0) \\ \delta X^{*R}(0) \end{bmatrix} \quad (66)$$

The next control $U^*(i+1)$ at the point $(t(i+1), X^*(i+1))$ where $t(i+1) = t(i) + \delta t^*(i)$, is computed by interpolation in 5-dimensional (t, X) -space. Here $\delta t^*(i)$ means $\delta t(U^*(i))$. The next time interval, state vector, and optimal value can be computed as above. The next block is to be transferred from disk memory to high speed memory, if the next state is a boundary point of this block.

These computations continue until the final condition is met. So the sequences of optimal controls, optimal values and the corresponding times are obtained.

REFERENCE

- [1] Athans, M. and Falb, P.L., Optimal Control: An Introduction to the Theory and its Applications, McGraw-Hill Book Company, New York, 1966
- [2] Bellman, R., Dynamic Programming, Princeton University Press, Princeton, N.J., 1957
- [3] Dolansky, L., "Present State of the Lanchester Theory of Combat," Opns. Res. Vol. 12, PP344-353, 1964
- [4] Dresher, M., Games of Strategy : Theory and Applications, Prentice-Hall, Inc., Englewood Cliffs, N.J., 1961
- [5] Finkbeiner, II, D.T., Introduction to Matrices and Linear Transformations, W.H. Freeman & Company, San Francisco, 1966
- [6] Helmer, O., Combat Between Heterogeneous Forces, RM-6, The Rand Corporation, 1947
- [7] Isaacs, R., Differential Games : A Mathematical Theory with Applications to Warfare and Pursuit, Control and Optimization, Robert E. Krieger Publishing Company, Huntington, New York, 1975
- [8] Isbell, J.R. and Marlow, W.H., "Methods of Mathematical Tactics," Logistics Papers, Issue No. 14, Logistics Res.

Proj., Contract N7 onr 41904, Proj. NRO47001, George Washington University, 1956

- [9] Kawara, Y., "An Allocation Problem of Support Fire in Combat as a Differential Game, Opns. Res., Vol. 21, PP942-951, 1973
- [10] Lang, S., Linear Algebra, Addison-Wesley Publishing Co., Reading, Massachusetts, 1970
- [11] Larson, R.E., "Dynamic Programming with Reduced Computational Requirements," IEEE Trans. on Automatic Control, Vol. Ac-10, PP135-143, 1965
- [12] _____, State Increment Dynamic Programming, American Elsevier Publishing Company, Inc., New York, 1968
- [13] Snow, R.N., Contributions to Lanchester Attrition Theory, RA-15078, The Rand Corporation, USAF Project MX-791, Douglas Aircraft Co. 1948
- [14] Taylor, J.G., "Target Selection in Lanchester Combat: Linear-law Attrition Process," Nav. Res. Log. Quart., Vol. 20, PP673-697, 1973
- [15] _____, "Lanchester-type Models of Warfare and Optimal Control," Nav. Res. Log. Quart., Vol. 21, PP79-106, 1974
- [16] _____, "Target Selection in Lanchester Combat: Heterogeneous Forces and Time-dependent Attrition-rate Coefficients," Nav. Res. Log. Quart. Vol 21, PP683-704, 1974
- [17] Von Neumann, J. and Morgenstern, O., Theory of Games & Economic Behavior, John Wiley & Sons, Inc., New York, 1944
- [18] Weiss, H.K., "Some Differential Games of Tactical Interest and the Value of a Supporting Weapon Systems," Opns. Res. Vol. 7, PP180-196, 1959

SOME EXPERIMENTS IN SEARCH THEORY

ALAN R. WASHBURN

Department of Operations Research
Naval Postgraduate School
Monterey, Ca. 93940, U.S.A.

ABSTRACT. The random search formula in its various forms enjoys wide application. This is remarkable for two reasons. Firstly, random search is in a formal sense impossible to carry out. Secondly, it represents a complete lack of organization that a searcher would presumably want to avoid, rather than emulate. This paper will summarize the results of some free play experiments involving live subjects. The experiments are designed to test whether the theory has any predictive power in situations where the target is able to maneuver in such a way as to avoid being detected.

1. INTRODUCTION

The theory of random search plays a central role in search theory. It often permits simple computation of the probability of detection without going into the details of exactly how the searcher and perhaps the target will or should move about during the search. However, the theory depends on an independence assumption that is in most cases impossible to fulfill in a strict sense, so that its predictions should be suspect without experimental verification. In this paper, the results of two experiments intended to verify some predictions of the theory will be recounted.

2. THE THEORY OF RANDOM SEARCH

A searcher begins to search for a target at time 0. We make two assumptions:

- 1) There is a "detection rate" function $\gamma(t)$ such that the probability of a detection in a small interval of time Δ that includes t is $\Delta\gamma(t)$.
- 2) The events that there is no detection in any set of non-overlapping intervals are independent.

These two assumptions are sufficient to determine the probability $q(t)$ that there will be no detection in the interval $[0, t)$. Since the events of no detection in the intervals $[0, t)$ and $[t, t + \Delta)$ are independent,

$$q(t + \Delta) = q(t)(1 - \Delta\gamma(t)) \quad \text{for small } \Delta \quad (1)$$

Rearranging equation (1),

$$[q(t + \Delta) - q(t)]/\Delta = -q(t)\gamma(t) \quad \text{for small } \Delta \quad (2)$$

Taking the limit as $\Delta \rightarrow 0$,

$$\frac{d}{dt} q(t) = -q(t)\gamma(t), \quad (3)$$

or

$$\frac{d}{dt} (\ln q(t)) = -\gamma(t). \quad (4)$$

Since $q(0) = 1$,

$$q(t) = \exp(-n(t)), \quad \text{where} \quad n(t) = \int_0^t \gamma(u) du. \quad (5)$$

Equation (5) is the general formula for random search. The number of detections up to time t is a non-homogeneous Poisson Process, with $n(t)$ being the mean and $q(t)$ being the probability of none.

Experiment 1: Expanding Area Search

The two-dimensional position of a target is accurately known at time 0, but for some reason search cannot begin until time τ , at which time the searcher begins searching with speed V and sweepwidth W . The target's maximum speed is U . The goal of the target is to evade detection. We wish to estimate the probability $p(t) = 1 - q(t)$ that the target will be detected at some time before t .

At time t , the target can be anywhere within a farthest-on circle with radius tU and area $A(t) = \pi t^2 U^2$. In an interval of time Δ , the searcher covers an area $B = VW\Delta$. If the searcher could search in such a manner that the successive incremental searched areas B were located independently of each other, but nonetheless all uniformly distributed within the farthest-on circle, then the random search formula with $\Delta\gamma(t) = B/A(t)$ would hold regardless of the strategy of the target. The search pattern alone would be sufficient to ensure that the two required assumptions hold, regardless of how the target's position depends on time. Similarly, if the target could maneuver in such a manner that his position were uniformly distributed within the farthest-on circle in every small time interval, but nonetheless independent of his position in every other time interval, then the random search formula with $\Delta\gamma(t) = B/A(t)$ would hold regardless of the strategy of the searcher. In other words, if the motions described above were feasible, then the random search formula would represent the value of the game.

Of course, neither of the above motions is feasible, since each player's position would have to hop about in a manner that is simply not characteristic of motion with a bounded speed. Nonetheless, since the random search formula would hold if either player could move as specified, it makes intuitive sense to expect that the formula will hold if both players move "as randomly as possible." In any case, this was the hypothesis that was investigated experimentally.

Since $\gamma(t) = VW/(\pi U^2 t^2)$ in this case,

$$n(t) = \int_{\tau}^t \gamma(u) du = (VW/\pi U^2)(1/\tau - 1/t) \quad \text{for } t \geq \tau. \quad (6)$$

Note that $n(\infty) = VW/\pi U^2 \tau$ is finite, which means that $q(\infty)$ is not 0. Since the area $A(t)$ expands quadratically with time, the job of the searcher eventually becomes hopeless; either the target will be found early, or not at all. The function $1 - q(t)$ is compared with experimental results in Figure 1.

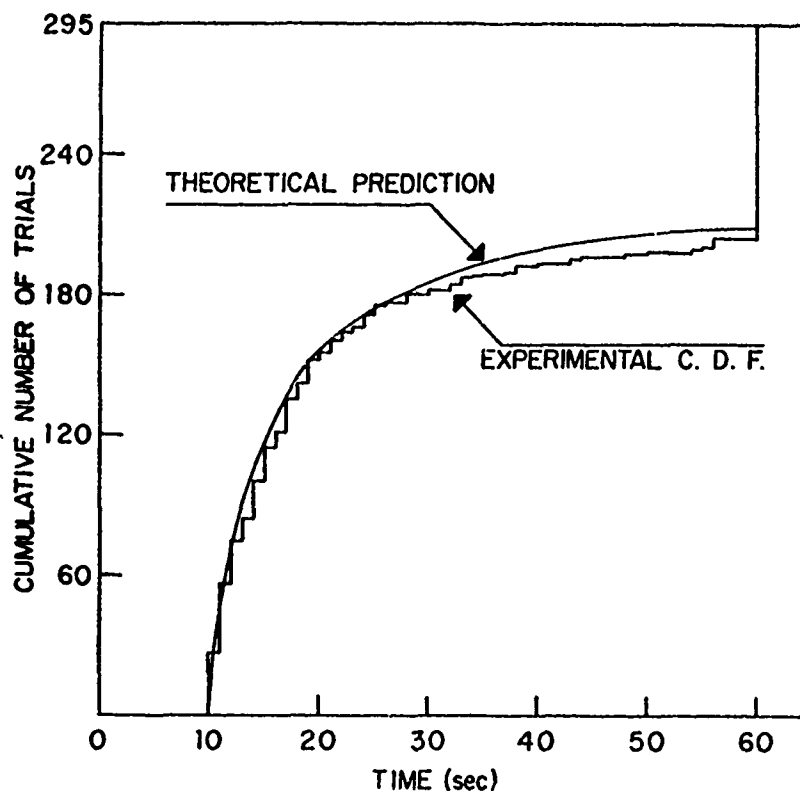


Figure 1:

V = PURSUER SPEED = .192 unit / sec
 U = EVADER SPEED = .024 unit / sec
 W = SWEEP WIDTH = .14 unit
 τ = TIME LATE = 10 sec.

The experiment was performed using officer-students at the United States Naval Postgraduate School as subjects, in pairs. Each player controlled his own position on his own cathode ray tube using a joystick, subject only to a velocity constraint. The displays were arranged in such a manner that

neither player could see the other's display. The continually expanding farthest-on circle was displayed on both displays. No instructions were given, except that both players were told that the searcher desired detection, whereas the target desired to avoid it. Play was terminated electronically when the searcher first came within $W/2$ of the target. The experimental cumulative distribution function for 295 trials is compared with $(1 - q(t))$ in Figure 1, showing the good agreement of theory with experiment.

Experiment 2: Fixed Area Search

In this experiment, the target must stay within a rectangle of area A , else he is counted as caught. The searcher has speed V and sweepwidth W , and starts at a random place within A . The target also starts at a random place within A , but otherwise a variety of assumptions are made:

- 2a) The target does not move. This experiment was not performed, since an exhaustive search is possible and presumably the time to detection would be uniform in $[0, A/VW]$. Amongst the various reasons usually given why an exhaustive search is often not possible in practice (poor navigation, non-cookie-cutter sensor, target motion, etc.), the simplest to simulate was target motion. So the first experiment to be performed was 2b.
- 2b) The target moves at speed $U = .2V$. The expectation was that the target would move randomly, that this motion would turn attempted exhaustive searches into random searches, that $\gamma(t)$ would therefore be VW/A , and $1 - q(t) = 1 - \exp(VWt/A)$. These expectations were verified. With $A/VW = 270$ sec, the mean time to detection in 38 trials was 265 seconds. The reason for the small number of trials is that this is an extremely boring game to play.
- 2c) Same as 2b) except that the target's display includes a cursor that always points toward the searcher. The target's motion should no longer be expected to be random, but one might argue that U is so small compared to V that the time to detection should still be exponential with a mean of 270 seconds, as in 2b). This turned out not to be the case. The mean time to detection in 131 trials was 367 seconds. However, the time to detection was still exponential, which is the same as saying that the detection rate function was a constant (see Figure 2).

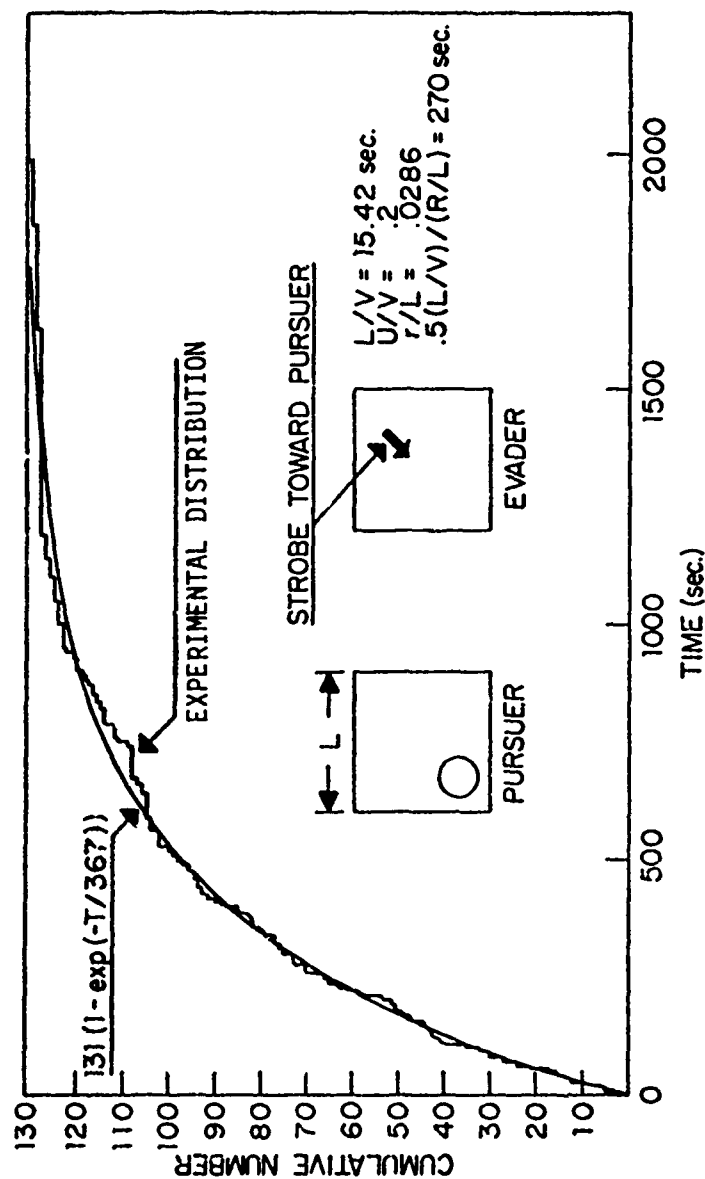


Figure 2: EVADER KNOWS PURSUER'S DIRECTION

2d) Same as 2b) except that the position of the searcher appears on the target's display. In this case the mean time to detection in 76 trials was 407 seconds. As in 2b) and 2c), the time to detection was an exponential random variable.

3. SUMMARY

The theory of random search, coupled with basic arguments about the detection rate function, has remarkably great but nonetheless imperfect powers of prediction. In the class of search problems where a target is lost within a fixed area, the hypothesis that the time to detection is an exponential random variable appears to be robust. In the expanding area search, the theory is accurate quantitatively, as well as qualitatively.

4. REFERENCE

- [1] Koopman, B. O., THE THEORY OF SEARCH II. TARGET DETECTION, Opns. Res., Vol. 4, PP. 503-531, 1956.

PARAMETRIC ANALYSIS OF MAIN
BATTLE TANK MOBILITY IN
KOREAN TERRAIN

ALAN S. THOMAS

WILLIAM A. NIEMEYER and ROBERT C. THIBODEAU

Combat Support Division
US Army Materiel Systems Analysis Activity
Aberdeen Proving Ground, MD 21005, U.S.A.

ABSTRACT. This report evaluates, through parametric component variation, the mobility potential of a series of main battle tank configurations operating in Korean terrain. Four principal design areas are varied, including the power train, suspension, weight and hull geometry. The speed potential of each vehicle configuration is determined throughout the terrain spectrum, and non-negotiable terrain is identified. Diagnostic analysis indicates the factors causing speed limitation as well as the reasons when the terrain is impassable. Some detail is provided on the soft soil mobility of the various configurations and the significance of differences on the basis of Korean seasonal soil strength.

1. INTRODUCTION

This report evaluates, through parametric component variation, the mobility potential of a series of main battle tank configurations operating in Korean terrain. Four principal design areas are varied, including the power train, suspension, weight and hull geometry. The Korean terrain as characterized by the Waterways Experiment Station, Vicksburg, Mississippi, is statistically described, and the methodology for modeling vehicle performance is summarized.

The speed potential of each vehicle configuration is determined throughout the terrain spectrum, and non-negotiable terrain is identified. Diagnostic analysis indicates the factors causing speed limitation as well as the reasons when the terrain is impassable. Some detail is provided on the soft soil mobility of the various configurations and the significance of differences on the basis of Korean seasonal soil strength.

2. MODEL DESCRIPTION

The computer simulation used to evaluate the vehicle/terrain interaction and predict mobility performance is the Army Mobility Model (AMM). This model considers vehicle performance in both areal and linear type terrain features. The areal mobility prediction part of the model (which is the only portion used in this evaluation) is shown schematically in Figure 1. The fundamental operation of this model is as follows. Detailed areal terrain data are collected from existing terrain data sources such as topographical maps, air photos, terrain studies, agricultural data and soil maps. Where possible these data sources are supplemented by actual field surveys. All these data sources are then used to develop a series of individual maps of the area being considered for each of the terrain factors shown in Figure 1.

The terrain input processor accepts these maps and overlays them to define areas in which the terrain is homogeneous with respect to all of the terrain factors simultaneously. The result of this process is an areal terrain unit map as shown, where unit number 98 might reflect an area where the slopes are uniformly between 5 and 10 percent and the soil strength in the wet season is uniformly between 40 and 60 cone index, etc. Associated with each map unit number is a range of values for each of 14 terrain factors. The factor categories are shown in Table 1.

The model requires a total of 76 vehicle characteristic inputs. These range from vehicle size and weight to details of its power train and suspension components. With these

TABLE 1 - TERRAIN CLASSIFICATION SYSTEM

TERRAIN FACTORS	1	2	3	4	5	6	7	8	9	10	11	12	13	14
Surface Type	Flag Flag Grained (other)	Coarse Grained	Medium	CH										
Surface strength (CI or NCI) Class range	201-300	221-200	161-220	101-160	61-100	41-60	33-40	26-32	17-25	11-16	4-10			
Slope (%) Class range	1-2	3-5	6-10	11-20	21-40	41-60	61-70	70-90						
Obstacle approach angle (deg) Class range														
Obstacle vertical ang (in) Class range	170	101	170-170	102-104	371-173	85-190	159-170	191-202	149-158	202-211	136-149	212-225	90-135	236-270
Obstacle base width (in) Class range	3-6	7-10	11-14	15-18	19-24	25-33	33-45	46-60						
Obstacle length (ft) Class range	40-150	34-47	24-35	22-24	6-12									
Obstacle spacing (ft) Class range	1	2-3	4-6	7-10	11-20	21-40	41-100							
Obstacle spacing type	100	66-107	57-65	26-36	19-25	13-18	9-12	2-8						
Surface roughness \pm 10 (ft in.)	1	2-4	5-6	7-8	9-12	13-16	17-22	23-32	33-45					
Stem diameter (in.) Factor value	0	1	2-4	5-8	9-12	13-16	17-22	23-32	33-45					
Stem spacing (ft)	320	64-327	34-65	26-35	18-25	13-17	8-12	1-7						
Visibility (ft) Class range	145-300	77-164	38-78	30-38	20-29	14-19	10-13	6-9	1-5					
Urban Code	Village	Town	City	Urban										

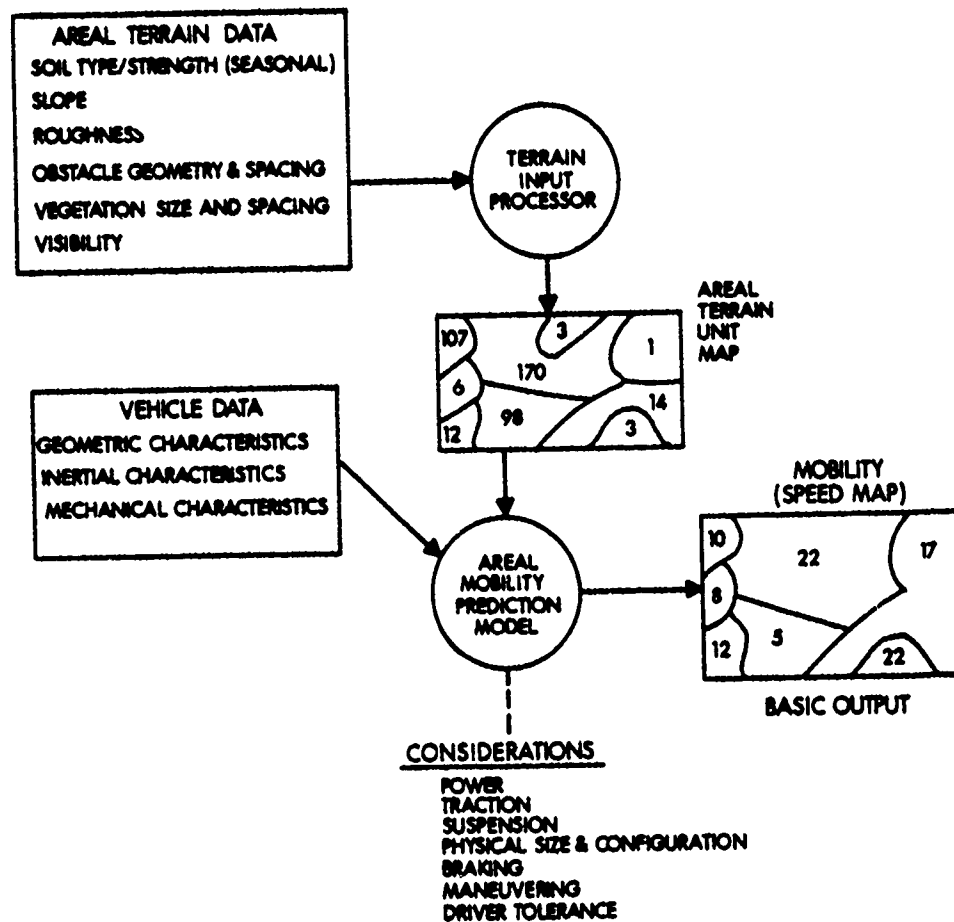


Figure 1. AMC Mobility Model (Areal Mobility Prediction)

data the various mathematical submodels of the overall model predict vehicle performance in the terrain factor values established for each map unit.

Submodels consider vehicle performance in the following manner:

<u>Terrain Factors Considered</u>	<u>Vehicle Performance Predicted</u>
Soil Type	Tractive and resistance forces throughout speed range.
Soil strength	
Slope	
Terrain roughness	Ride limited speed
Obstacles	Hangup, traction, dynamic loading, acceleration and braking between obstacles.
Vegetation	Traction for overriding, and vehicle size for maneuvering between trees. Driver visibility.

For a given map unit the speed results of each of these submodels are compared for uphill, downhill, and level slope conditions; the limiting value is selected for each condition, and the three limiting values are averaged to provide the vehicle's estimated best speed in that map unit. In considering the vegetation factor the model examines various strategies of maneuvering around certain size trees and overriding others to obtain the best vehicle speed. Some terrain factors such as soil strength and slope naturally interact with others, so are considered simultaneously. For example, a vehicle on a soft soil slope will have less tractive force available to climb an obstacle or override a tree than it would on a level hard surface because some of its tractive force capability is used in overcoming the soft soil motion resistance and the grade resistance. The basic speed output of the model can be used to develop a speed map as shown in Figure 1.

3. TERRAIN DESCRIPTION

The terrain used in this analysis is an area approximately 12 kilometers wide and 46 kilometers long on the northeast coast, between 41° and 42° latitude. The terrain was characterized, in accordance with Table 1, by the US Army Corp of Engineers, Waterways Experiment Station. It is the only Korean terrain that has been so characterized, and it is

not known to what degree it can be considered representative of that in South Korea. This terrain was originally chosen for characterization simply because there was a complete set of topographic maps available for this area and it appeared to be reasonably representative of an area in which vehicle negotiation would be practical. The subject area is shown schematically in Figure 2 and the map sheet locations are shown in Figure 3.

The area has been divided into approximately 2,000 discrete terrain units, but several of the non-contiguous units have identical characteristics, so that there are 1617 unique combinations of terrain factors, i.e., there are 1617 different types of terrain units that occur. Appendix A contains frequency distributions (as represented by percent area rather than number of terrain units) for the most significant factors of the terrain characterization. The area is shown to be very hilly with steep slopes, but at the same time there is a significant portion of the terrain with soil too soft to support repeated traffic of an M60. There is also a significant portion of the terrain with high surface roughness. In virtually all respects, the Korean terrain analyzed herein is more severe than the West German terrain typically used to represent European operations.

4. VEHICLE CONFIGURATIONS - PARAMETRIC VARIATIONS

There are four principal design areas that are addressed parametrically in this report. These are the power train, suspension, gross weight, and general hull configuration including both geometric and inertial characteristics. Each of these factors are evaluated at two different levels with the exception of the power train which is addressed at three performance levels. The base level in all cases represents the current M60 series design. The upper level represents the practical "state of the art" in main battle tank design and will be referred to as the "SOTA" level. The third power train level is intermediate to the upper and lower levels and is believed to represent the performance available by product improvement of the M60 engine without a transmission change.

The result of providing the variation discussed above is a matrix yielding 24 vehicle configurations. These are identified in Table 2. It may be noted that the weight levels selected are not entirely consistent with the M60/SOTA spectrum of vehicle characteristics. The lower weight level is especially artificial, but was nonetheless selected after discussions with visitors from the Republic of Korea and TARADCOM in April, 1977.

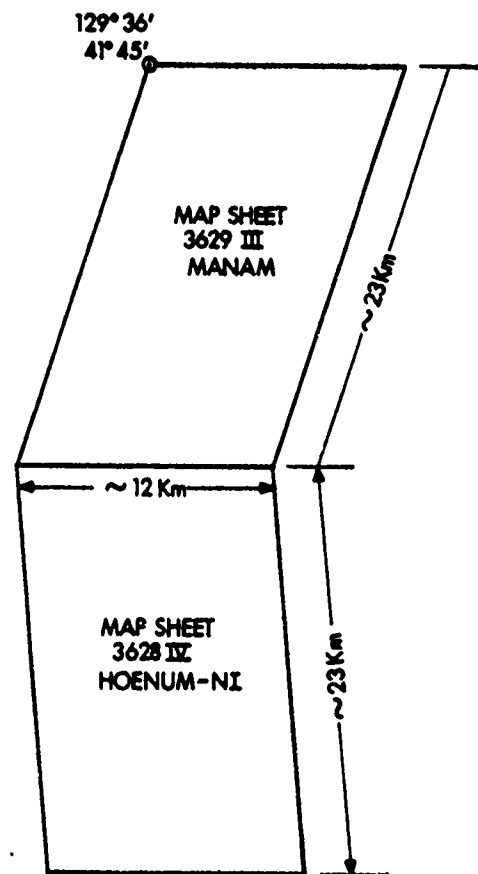


Figure 2. Study Terrain Map Sheets.

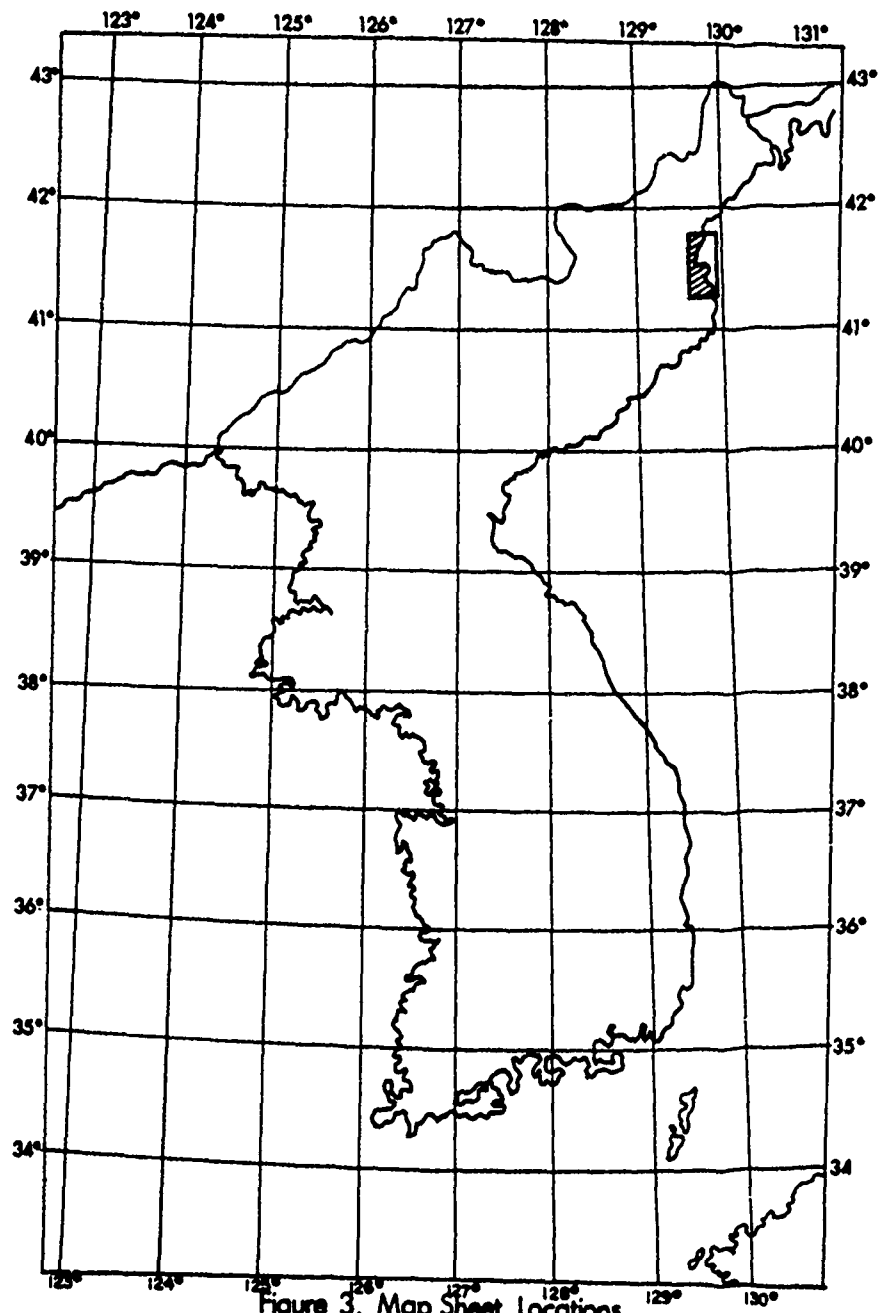


Figure 3. Map Sheet Locations.

TABLE 2 . VEHICLE CONFIGURATION MATRIX

Configuration Number	Power Train (HP)	Suspension	Weight (TONS)	Hull Configuration
1	750	M60	45	M60
2	750	M60	55	M60
3	900	M60	45	M60
4	900	M60	55	M60
5	1500	M60	45	M60
6	1500	M60	55	M60
7	750	SOTA	45	SOTA
8	750	SOTA	55	SOTA
9	900	SOTA	45	SOTA
10	900	SOTA	55	SOTA
11	1500	SOTA	45	SOTA
12	1500	SOTA	55	SOTA
13	750	SOTA	45	M60
14	750	SOTA	55	M60
15	900	SOTA	45	M60
16	900	SOTA	55	M60
17	1500	SOTA	45	M60
18	1500	SOTA	55	M60
19	750	M60	45	SOTA
20	750	M60	55	SOTA
21	900	M60	45	SOTA
22	900	M60	55	SOTA
23	1500	M60	45	SOTA
24	1500	M60	55	SOTA

5. MODELING RESULTS

The mobility model described in Section II was utilized to predict the performance of each vehicle configuration across the 1617 terrain unit types identified in the Korean terrain. The results of these analyses are presented in the form of mobility profiles in Figures 4 through 11. The profiles in Figures 4 through 7 are generated by ordering the terrain units along the horizontal axis with those providing best vehicle performance considered first (furthest to the left on this axis). The vehicle speed in each individual terrain unit is then plotted as a function of the terrain unit position on the area axis (which, as stated, is determined by trafficability). Thus the actual vehicle speed which can be obtained in, for example, the 75th percentile terrain, is depicted. Figures 8 through 11 use the same technique for arranging the terrain units, but rather than plot the actual speed in each unit, the cumulative average speed over all terrain units to the left of each point on the horizontal axis is plotted. Thus from these curves one can determine the average speed of the vehicle if it is driven in, for example, the most trafficable 75 percent of the terrain, with the most severe 25 percent avoided. One further note on computational techniques - when the model determines a terrain unit to be impassable, it nevertheless assigns a vehicle speed of 0.1 mph to that unit as a computational expediency. Thus on the cumulative average speed curves, the 0.1 mph is averaged with all prior unit speeds and a finite average speed is predicted even over impassable terrain. However, reference to the actual speed curves will indicate the point at which this artificiality occurs.

In order to summarize the results shown in preceding mobility profiles, Table 3 has been included. In this table, the average speed attainable with each configuration is shown, first when operating in the most trafficable 50% of the terrain (V_{50}), and also when in the most trafficable 80% of the terrain (V_{80}). The V_{50} performance might be representative of movement potential where the unit has not been forced to tactical deployment and route selection is not restricted. The V_{80} performance might then be taken as an indicator of movement potential under tactical deployment where route selection is more restricted. These are admittedly arbitrary measures, but nevertheless are believed to provide a reasonable basis for quantifying the effects of the parametric variation of vehicle components. The final quantification of these effects is obtained by determining the average contribution across all configurations, afforded by variation of a single vehicle

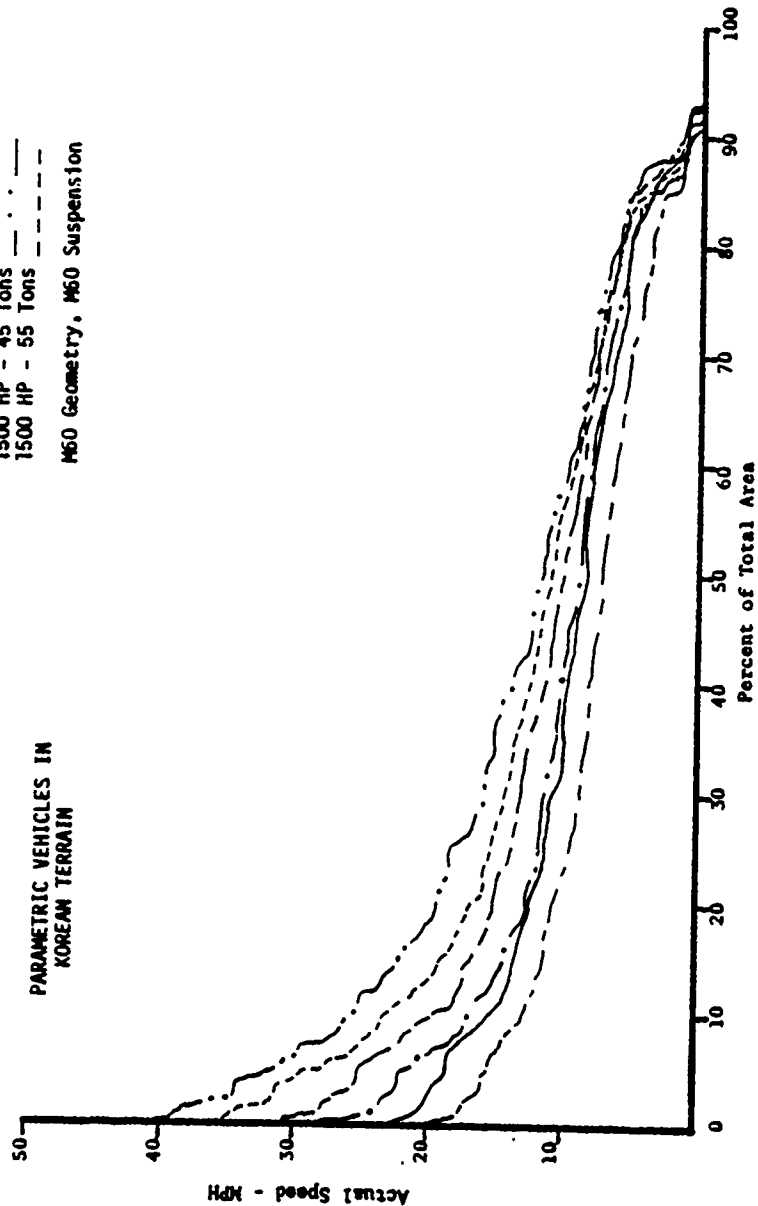
HORSEPOWER-WEIGHT GUIDE

750 HP - 45 Tons	---
750 HP - 55 Tons	---
900 HP - 45 Tons	---
900 HP - 55 Tons	---
1500 HP - 45 Tons	---
1500 HP - 55 Tons	---

M60 Geometry, M60 Suspension

Figure 4:

PARAMETRIC VEHICLES IN
KOREAN TERRAIN



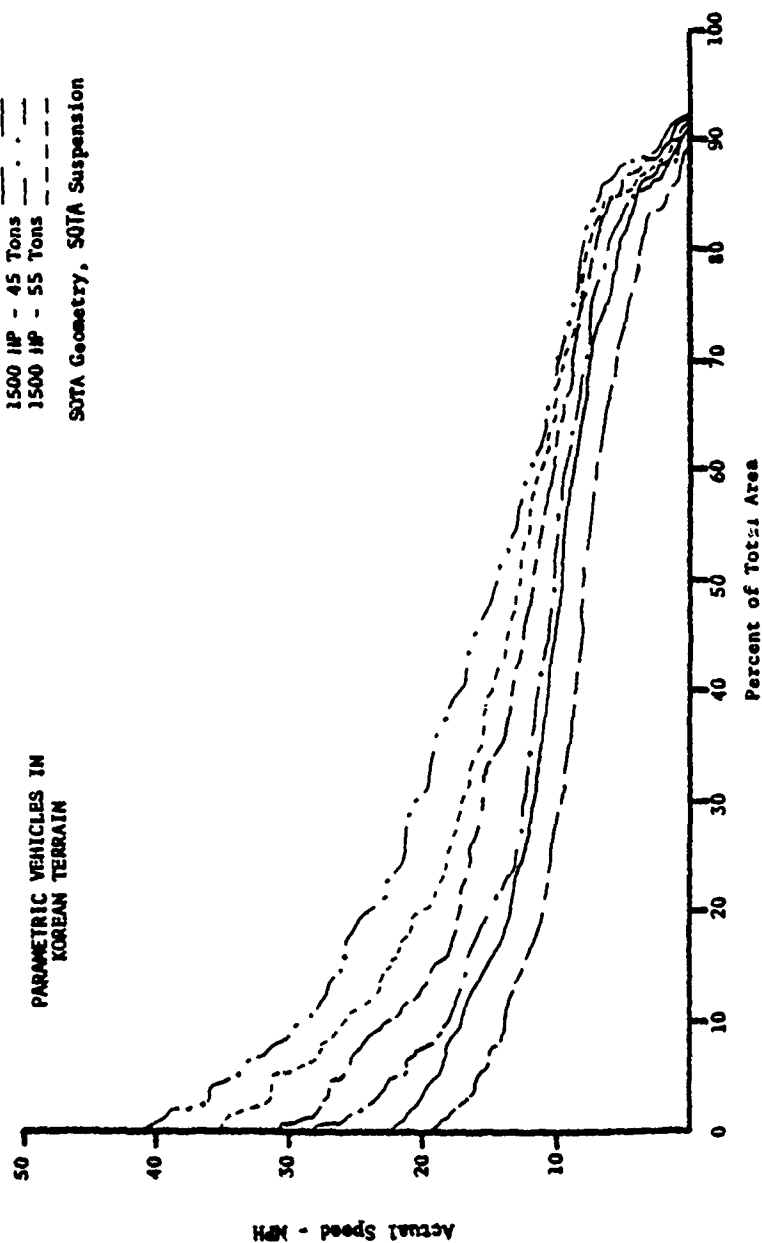
HORSEPOWER-WEIGHT GUIDE

750 HP - 45 Tons	-----
750 HP - 55 Tons	-----
900 HP - 45 Tons	-----
900 HP - 55 Tons	-----
1500 HP - 45 Tons	-----
1500 HP - 55 Tons	-----

SOTA Geometry, SOTA Suspension

Figure 5:

PARAMETRIC VEHICLES IN
KOREAN TERRAIN

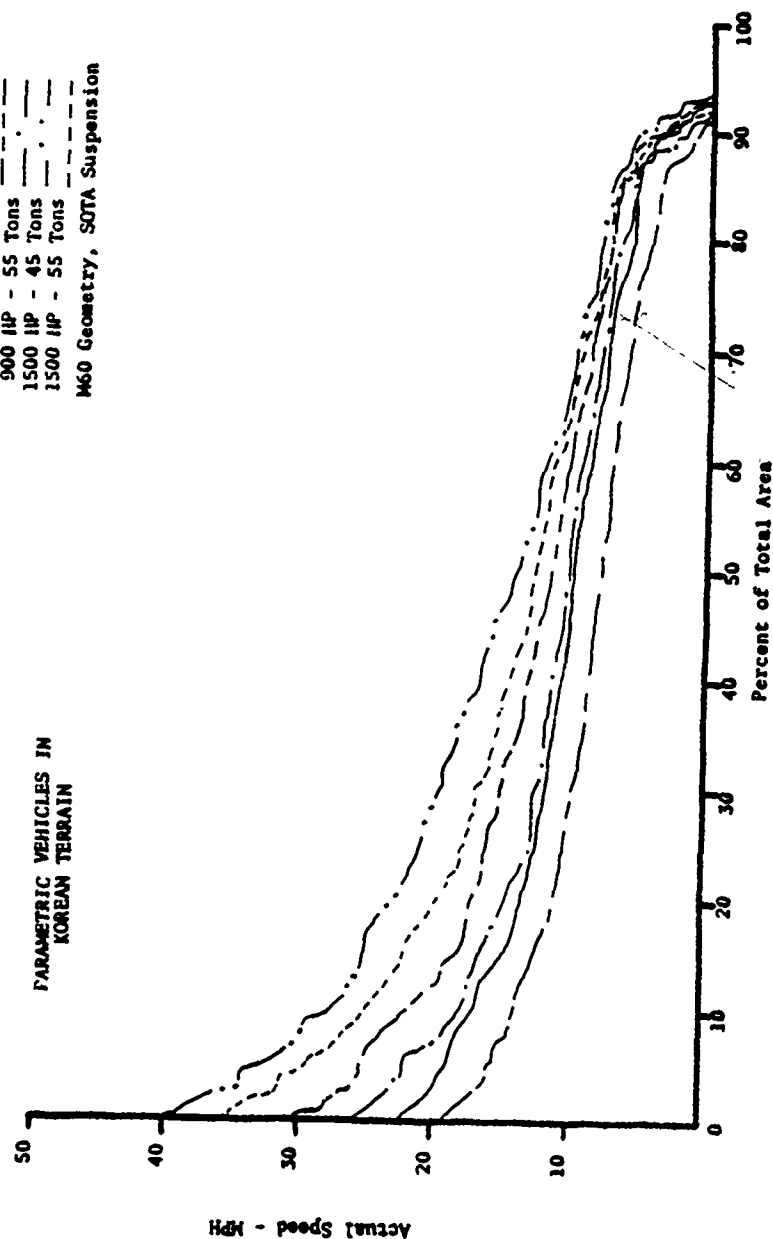


HORSEPOWER-WEIGHT GUIDE

750 HP	- 45 Tons
750 HP	- 55 Tons
900 HP	- 45 Tons
900 HP	- 55 Tons
1500 HP	- 45 Tons
1500 HP	- 55 Tons
M60 Geometry, SOTA Suspension	

Figure 6:

PARAMETRIC VEHICLES IN
KOREAN TERRAIN



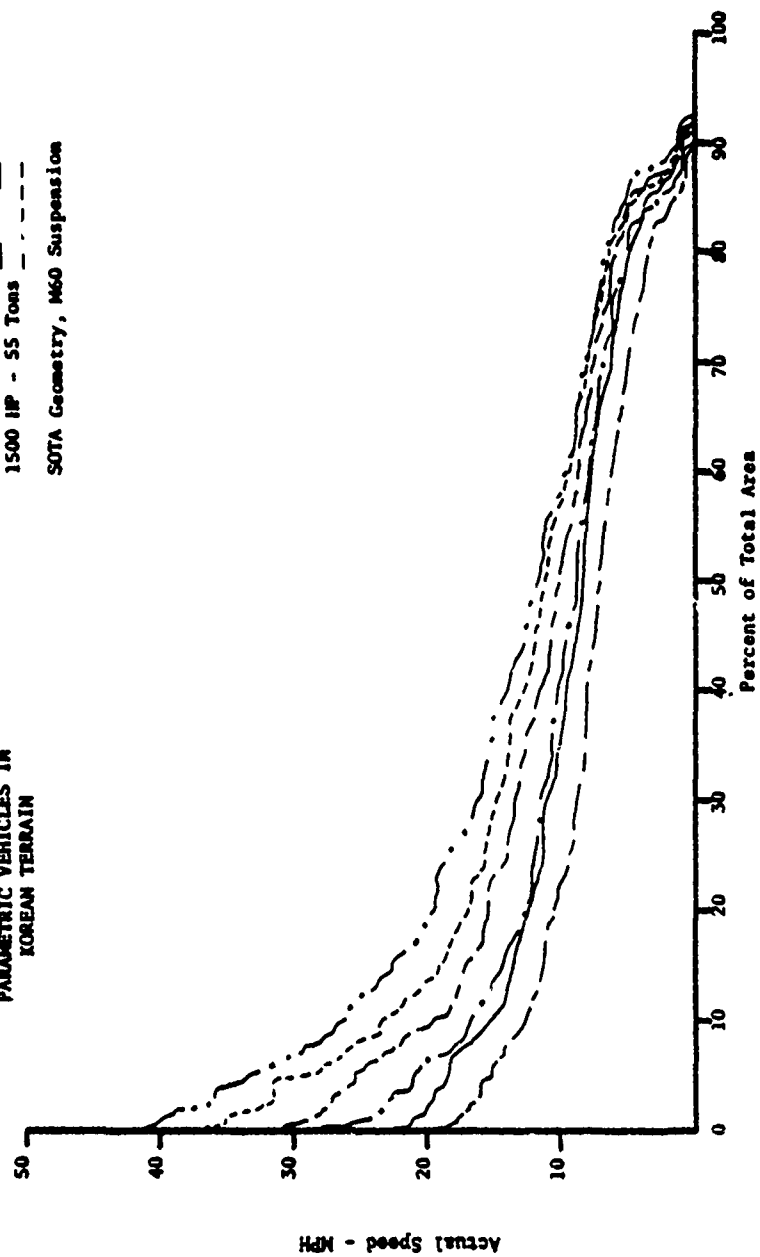
HORSEPOWER-WEIGHT GUIDE

750 HP - 45 Tons	_____
750 HP - 55 Tons	_____
900 HP - 45 Tons	_____
900 HP - 55 Tons	_____
1500 HP - 45 Tons	_____
1500 HP - 55 Tons	_____

SOTA Geometry, M60 Suspension

Figure 7:

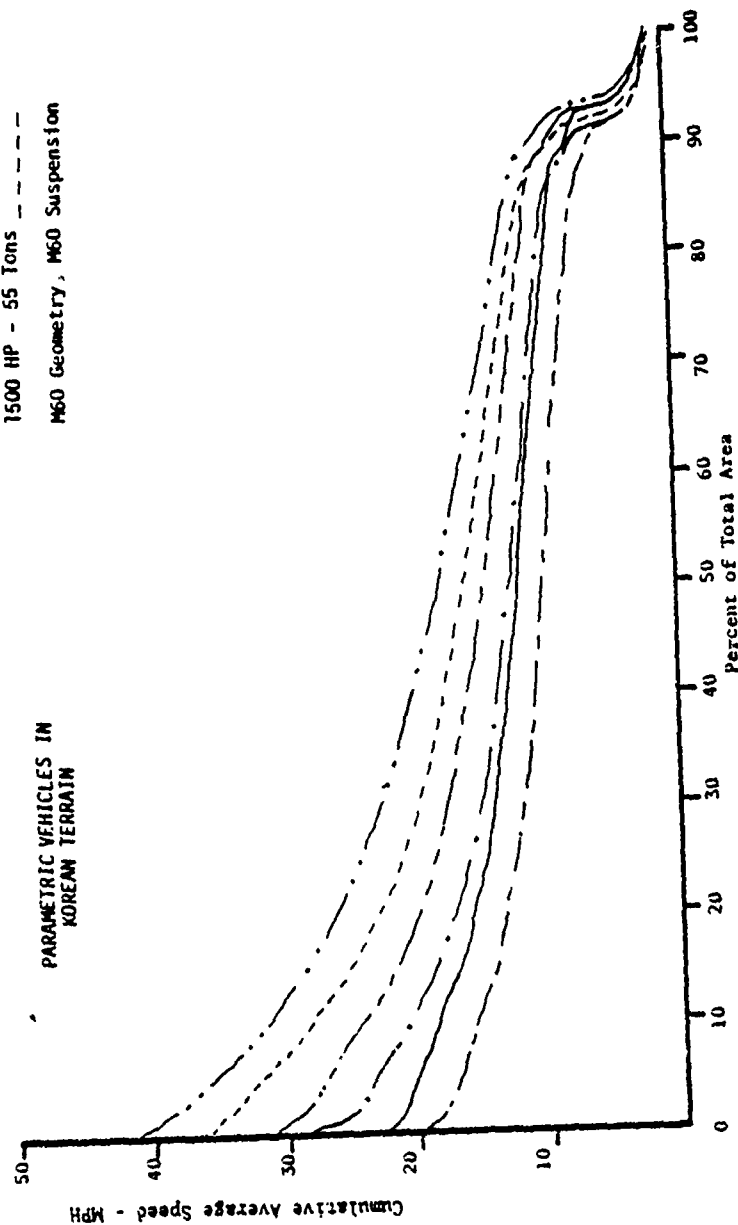
PARAMETRIC VEHICLES IN
KOREAN TERRAIN



HORSEPOWER-WEIGHT GUIDE

750 HP - 45 Tons	-----
750 HP - 55 Tons	-----
900 HP - 45 Tons	-----
900 HP - 55 Tons	-----
1500 HP - 45 Tons	-----
1500 HP - 55 Tons	-----
M60 Geometry, M60 Suspension	-----

Figure 8:

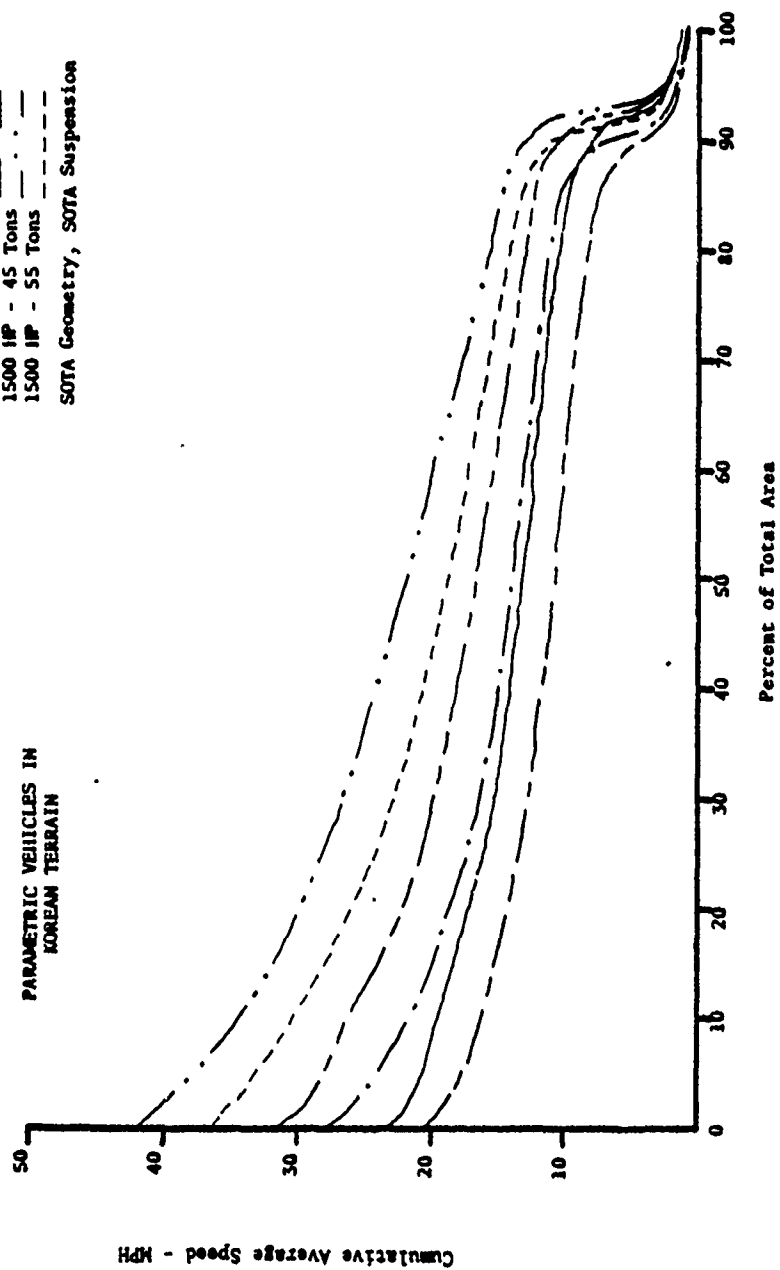


HORSEPOWER-WEIGHT GUIDE

750 HP	- 45 Tons
750 HP	- 55 Tons
900 HP	- 45 Tons
900 HP	- 55 Tons
1500 HP	- 45 Tons
1500 HP	- 55 Tons

SOTA Geometry, SOTA Suspension

Figure 9:
PARAMETRIC VEHICLES IN
KOREAN TERRAIN



IONSEPOWER-WEIGHT GUIDE

750 HP - 45 Tons	_____
750 HP - 55 Tons	_____
900 HP - 45 Tons	_____
900 HP - 55 Tons	_____
1500 HP - 45 Tons	_____
1500 HP - 55 Tons	_____

M60 Geometry, SOTA Suspension

Figure 10:

PARAMETRIC VE/ ICLES IN
KOREAN TERRAIN

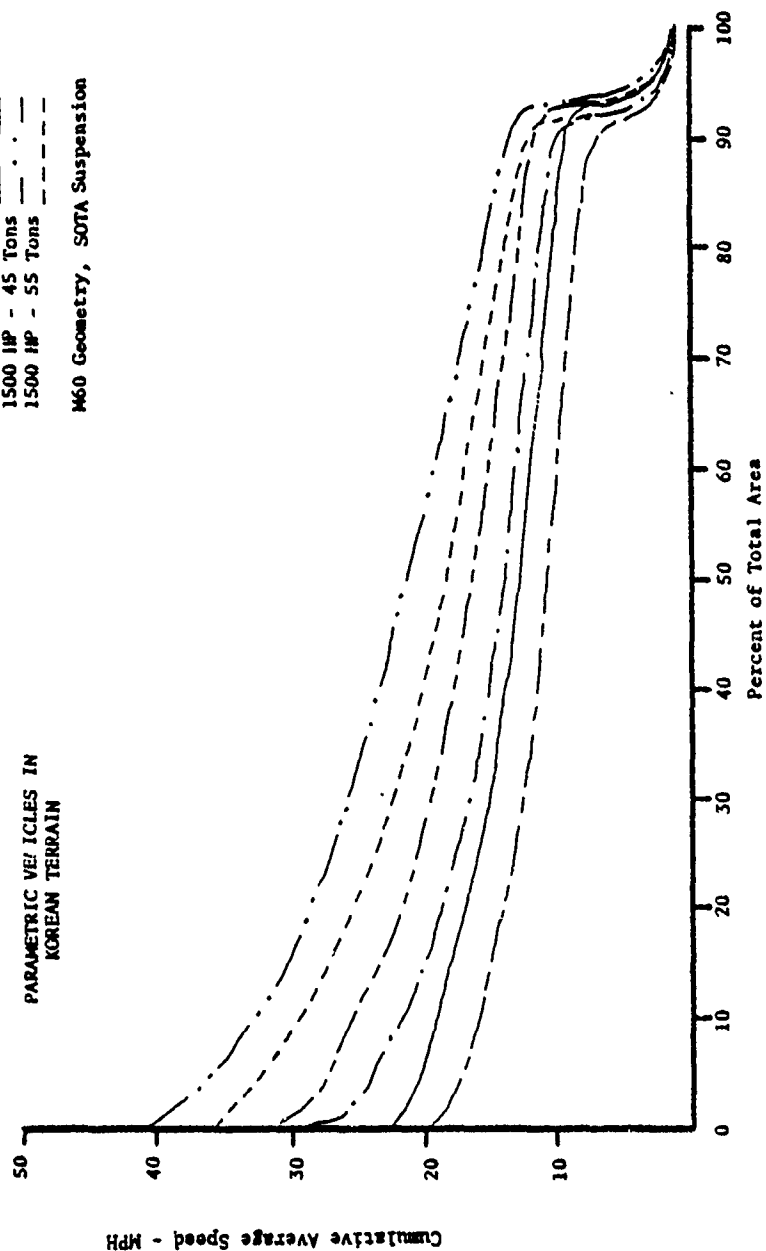


Figure 11:

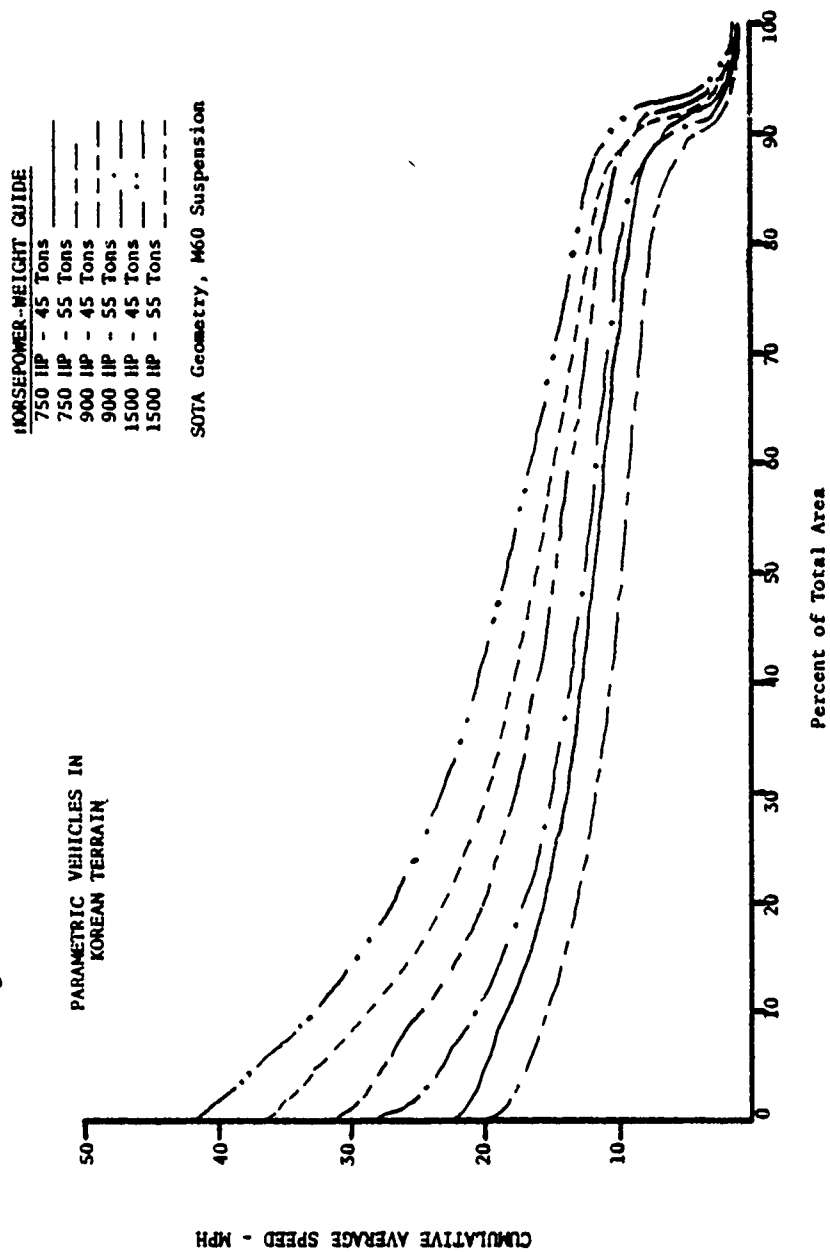


TABLE 3. SUMMARIZED CROSS COUNTRY SPEED PERFORMANCE

Conf. (GEOM-SUSP-PWR-WT) No.	Cumulative Average Speed (MPH)	
	V 50	V 80
1 (M60-M60-750-45)	11.7	9.2
2 (M60-M60-750-55)	9.9	7.5
3 (M60-M60-900-45)	14.5	11.1
4 (M60-M60-900-55)	12.5	9.8
5 (M60-M60-1500-45)	17.9	12.9
6 (M60-M60-1500-55)	16.0	11.9
7 (SOTA-SOTA-750-45)	12.8	10.2
8 (SOTA-SOTA-750-55)	10.6	8.1
9 (SOTA-SOTA-900-45)	16.6	12.8
10 (SOTA-SOTA-900-55)	13.8	11.0
11 (SOTA-SOTA-1500-45)	21.7	15.6
12 (SOTA-SOTA-1500-55)	18.6	14.1
13 (M60-SOTA-750-45)	12.9	10.3
14 (M60-SOTA-750-55)	10.7	8.5
15 (M60-SOTA-900-45)	16.4	12.8
16 (M60-SOTA-900-55)	13.8	11.1
17 (M60-SOTA-1500-45)	21.2	15.4
18 (M60-SOTA-1500-55)	18.4	13.8
19 (SOTA-M60-750-45)	11.7	9.1
20 (SOTA-M60-750-55)	9.8	7.4
21 (SOTA-M60-900-45)	14.6	11.1
22 (SOTA-M60-900-55)	12.5	9.7
23 (SOTA-M60-1500-45)	18.3	13.0
24 (SOTA-M60-1500-55)	16.2	12.0

parameter. This is accomplished by computing the average percentage improvement in V_{50} and V_{80} when a single vehicle variable is changed from its lower performance level to the higher level (e.g. from M60 level to SOTA level, or from 55T to 45T). The results are shown in Table 4.

In addition to the speed profiles discussed above, one further product of the mobility modeling is a determination of the factor for each terrain unit that either causes the unit to be impassable, or provides the ultimate speed constraint.

Factor categories 1-4 below indicate impassability; categories 5-10 are speed constraints.

1. Soil strength insufficient
2. Available traction less than soil and slope resistance
3. Obstacle interference
4. Available traction less than total resistance (including vegetation and obstacle override).
5. Ride dynamics
6. Soil and slope resistance
7. Visibility
8. Maneuvering (through vegetation and obstacles)
9. Total resistance to movement (including vegetation and obstacle override).
10. Acceleration and deceleration between obstacles.

Table 5 indicates the percentage of the total terrain area for which each of the 10 factors was the operative constraint, for each vehicle configuration. For example, if one were interested in power train effects on speed, configurations 8 and 12 might be compared. Speed limiting factors numbered 6 and 9 deal with power train constraints, and by summing these frequencies together we see that an increase from 750 HP to 1500 HP decreases the area in which we are power train limited from 50.4% of the total area to 29.9%. (Power effects are present in some of the other limiting factors, but 6 and 9 are the key factors.) It should be noted that this table only indicates the frequency with which the various factors are constraining, but not the performance level at which the constraint occurs.

6. SOFT SOIL MOBILITY CONSIDERATIONS

In addition to the comprehensive analysis of cross-country mobility provided in the previous section, there is also an issue of vehicle performance in marginal soft soil to be addressed. The primary design parameters involved in marginal soft soil performance are the length and width of the track, the contact area of the track shoe, and the weight of the vehicle. Again the analysis was conducted with a high and a

TABLE 4. AVERAGE PERFORMANCE IMPROVEMENTS OBTAINED
FROM VEHICLE COMPONENT VARIATION

<u>Component</u>	<u>Variation</u>	<u>Improvement</u>	
		<u>V₅₀</u>	<u>V₈₀</u>
• Suspension	M60 → SOTA	12.5%	14.7%
• Power Train			
	750 HP → 900 HP	27.3%	27.6%
	900 HP → 1500 HP	29.3%	21.6%
• Hull Geometry	M60 → SOTA	0.5%	-0.4%
• Weight	55T → 45T	17.3%	15.9%

TABLE 5 Frequency of Limiting Factor Occurrence

CONF NO.	(GEOM-SUSP-PWR-MT)	LIMITING FACTOR NUMBER									
		1	2	3	4	5	6	7	8	9	10
1	(M60-M60-750-45)	2.7	0.4	1.0	2.9	17.7	9.1	3.1	15.7	24.9	22.5
2	(M60-M60-750-55)	4.5	0.1	0.8	3.0	15.8	9.9	2.7	15.3	27.4	20.7
3	(M60-M60-900-45)	2.7	0.4	1.0	2.9	16.8	9.3	8.0	15.9	18.9	24.1
4	(M60-M60-900-55)	4.5	0.1	1.0	2.9	15.8	9.4	6.6	15.8	21.7	22.2
5	(M60-M60-1500-45)	2.7	0.4	0.6	2.9	20.3	5.6	12.2	18.1	12.0	25.3
6	(M60-M60-1500-55)	4.5	0.1	0.6	3.0	19.2	7.1	10.0	16.9	15.2	23.7
7	(SOTA-SOTA-750-45)	3.1	0.1	2.2	2.9	8.6	13.6	3.1	16.9	34.9	14.7
8	(SOTA-SOTA-750-55)	4.6	0.0	2.8	2.9	7.9	14.4	2.5	16.1	36.0	12.9
9	(SOTA-SOTA-900-45)	3.1	0.1	1.7	2.9	11.2	11.1	10.0	17.2	26.5	16.3
10	(SOTA-SOTA-900-55)	4.6	0.0	2.2	2.9	9.9	11.3	8.1	16.6	29.9	14.7
11	(SOTA-SOTA-1500-45)	3.1	0.1	1.2	2.9	16.1	8.5	14.0	17.5	19.0	17.7
12	(SOTA-SOTA-1500-55)	4.6	0.0	1.3	2.9	14.7	9.0	11.8	17.8	20.9	17.1
13	(M60-SOTA-750-45)	2.7	0.4	0.8	2.9	8.4	12.4	5.4	13.7	34.3	19.1
14	(M60-SOTA-750-55)	4.5	0.1	0.8	3.0	7.8	12.9	4.6	12.5	36.1	17.8
15	(M60-SOTA-900-45)	2.7	0.4	0.7	2.9	10.7	9.7	13.4	14.0	25.2	20.3
16	(M60-SOTA-900-55)	4.5	0.1	0.8	2.9	9.6	9.8	10.9	13.7	28.8	19.1
17	(M60-SOTA-1500-45)	2.7	0.4	0.4	2.9	14.1	6.4	19.3	15.0	15.9	23.0
18	(M60-SOTA-1500-55)	4.5	0.1	0.4	3.0	13.3	7.6	16.0	14.9	19.2	21.2
19	(SOTA-M60-750-45)	3.1	0.1	2.4	2.9	18.6	9.2	1.8	17.9	25.0	19.0
20	(SOTA-M60-750-55)	4.6	0.0	2.8	2.9	16.6	10.4	1.3	17.5	27.2	16.8
21	(SOTA-M60-900-45)	3.1	0.1	2.0	2.9	17.1	10.5	5.9	17.9	20.7	20.0
22	(SOTA-M60-900-55)	4.6	0.0	2.4	2.9	15.9	10.8	4.9	17.4	22.2	19.0
23	(SOTA-M60-1500-45)	3.1	0.1	1.4	2.9	21.8	7.5	8.2	20.0	13.6	21.5
24	(SOTA-M60-1500-55)	4.6	0.0	1.5	2.9	20.2	8.4	6.8	18.9	16.1	20.5

lower level assigned to each of these factors so that the relative contribution of each might be identified. Table 6 identifies the levels selected for analysis of the track variables. It also shows the nominal ground pressure associated with each combination of variables at the two previously selected gross vehicle weights. Finally, the table shows a VCI_1 and VCI_{50} entry for each configuration. This is the minimum soil strength as measured by cone penetrometer readings which will permit one and fifty passes respectively, of the vehicle through the soil. The equations for calculating vehicle cone index (VCI) numbers have been empirically derived and are shown in Appendix B.

There is no direct relationship between vehicle cone index and ground pressure. However, over the relatively small variation in track parameters considered here, reasonably linear relationships can be established, as shown in Figure 12. The data points are those from Table 6. These trend lines provide a gross estimate of the soil strength required to support a vehicle with a given ground pressure.

In order to assess the significance of differences in ground pressure over the range analyzed (from about 9 psi to about 14 psi) it is necessary to determine the frequency with which soil strengths in the Korean terrain are sufficiently low as to influence vehicle passage. Figure 13 shows the cumulative frequency distributions of soil strengths for various seasons. The wet season and dry season distributions are obtained by referring to yearly average rainfall data and selecting the wettest and driest 30 day periods for analysis. The average season reflects soil conditions over the remaining 305 days. The fourth distribution (referred to as "WWET") is an estimate of conditions in severe periods of rainfall. It reflects the soil strength conditions over the 10 wettest days of the year but with the yearly averages for those days increased by 50 percent, as might be experienced in a particularly wet year.

Superimposed on the distribution curves in Figure 13 are the soil strengths required for both one and 50 vehicle passes at 9 and 14 PSI ground pressure. It may be observed that the significance of ground pressure is minimal insofar as the probability of completing one vehicle pass through Korean soft soil is concerned under any seasonal condition. However, the probability that the soil will support repeated traffic is very much affected by ground pressure. For example, in the wet season, about eight percent of the terrain will not support 50 passes of vehicles with 9 PSI ground pressure while 40% of the terrain will not support 50 passes if the vehicle ground pressure is 14 PSI.

Table 6 MARGINAL SOFT SOIL MOBILITY FACTORS

TRACK LENGTH (IN.)	TRACK WIDTH (IN.)	SHOE AREA (IN.) ²	45 TONS				55 TONS			
			NOM. GRD. PRESS. (PSI)	VCI ₁	VCI ₅₀	NOM. GRD. PRESS. (PSI)	VCI ₁	VCI ₅₀		
160	24	160	11.7	20.7	48.3	14.3	28.9	66.1		
160	24	180	11.7	20.6	48.0	14.3	28.8	65.8		
160	28	160	10.0	17.0	40.0	12.2	23.1	53.5		
160	28	180	10.0	16.8	39.8	12.2	23.0	53.2		
180	24	160	10.4	19.2	44.9	12.7	26.5	60.8		
180	24	180	10.4	19.1	44.6	12.7	26.4	60.5		
180	28	160	8.9	15.8	37.4	10.9	21.3	49.6		
180	28	180	8.9	15.7	37.2	10.9	21.2	49.3		

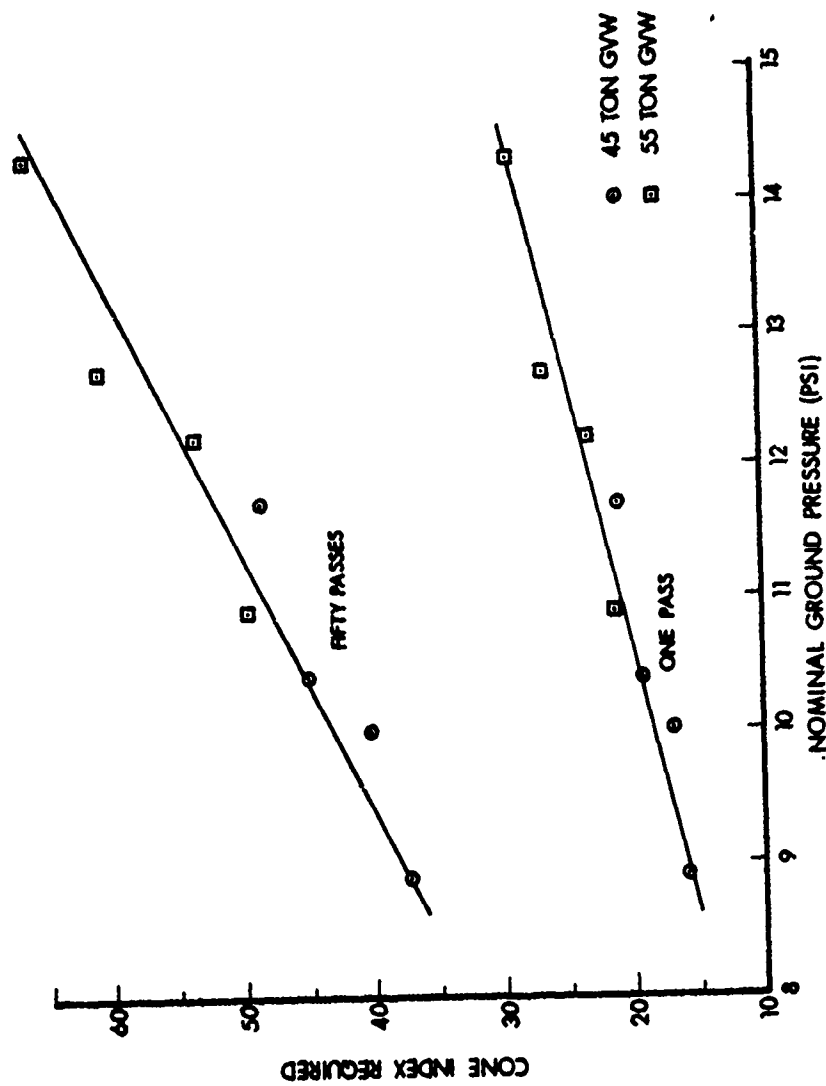


Figure 12. Gross Relationship Between Vehicle Ground Pressure and Soil Strength Required for Mobility

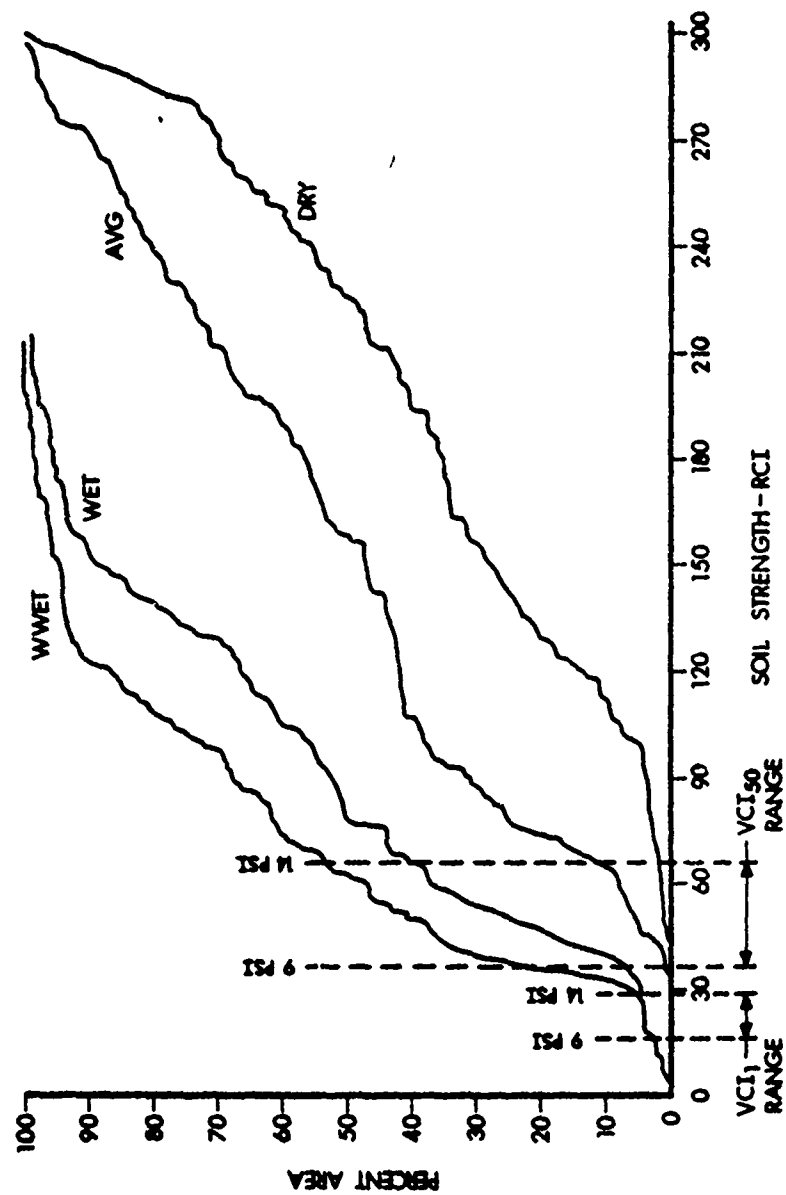


Figure 13. Soft Soil Limitation in Korean Terrain.

7. CONCLUSIONS

Keeping in mind that the results of the preceding section have been derived from analysis in the sample Korean terrain only, and assuming the parameter ranges used represent the feasible alternatives, the following conclusions are offered:

- Power train selection has the most critical impact on mobility potential. However, it is a case of diminishing returns. A relatively small increase from 750 HP to 900 HP produces about the same percentage improvement as does the much larger step in going from 900 HP to 1500 HP. (In both cases the absolute improvement in V_{80} is about 2.4 MPH.)

- The next best place to look for improved mobility is probably the suspension. Over the ranges considered the weight has a slightly greater effect, but weight reduction may mean a tradeoff in protection level. A suspension improvement produces about the same effect on mobility, but is more readily achieved.

- Movement rate is shown to be insensitive to hull geometry, however, lack of attention to this design factor can result in serious obstacle interference problems.

- Based on the above conclusions and using V_{80} as an index of mobility performance, Figure 14 shows the incremental improvements afforded by component changes ordered according to their payoff.

- Finally, with regard to soft soil mobility, it appears that the criticality of the track configuration is very dependent on gross vehicle weight. A 45-ton vehicle will have few trafficability problems with most of the track configurations. For areas of repeated traffic, it is desirable to maintain a ground pressure of 11 PSI or less. For a 55-ton vehicle the acceptable track configurations are thus narrowed considerably.

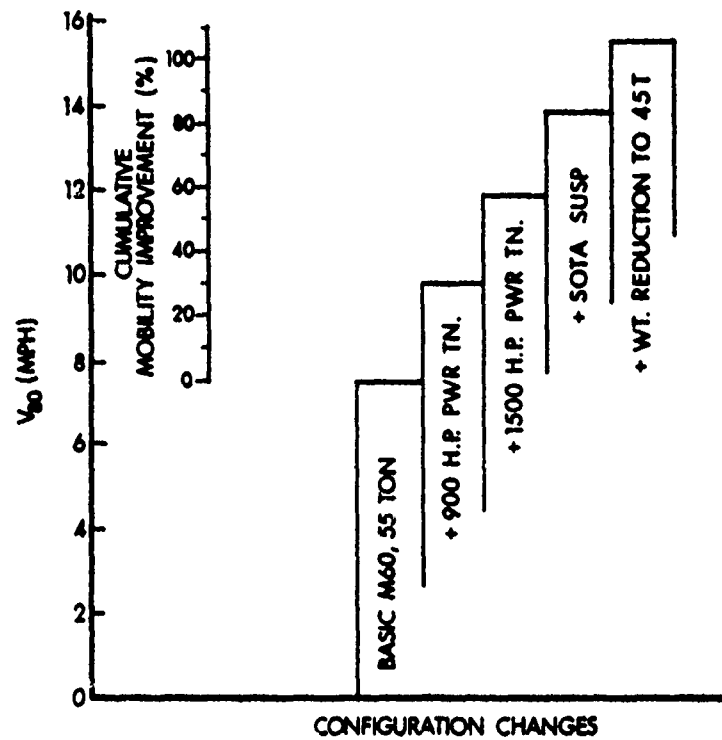


Figure 14. Mobility Enhancement Provided by Configuration Changes.

APPENDIX A

TERRAIN FACTOR DISTRIBUTIONS

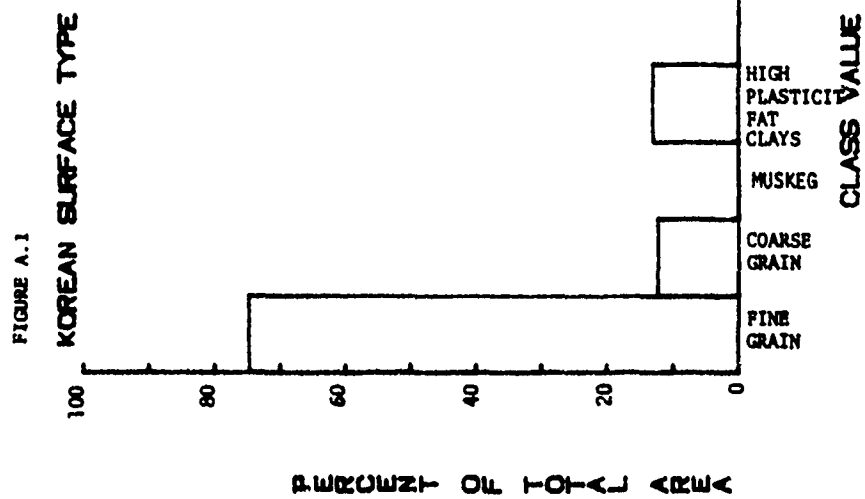
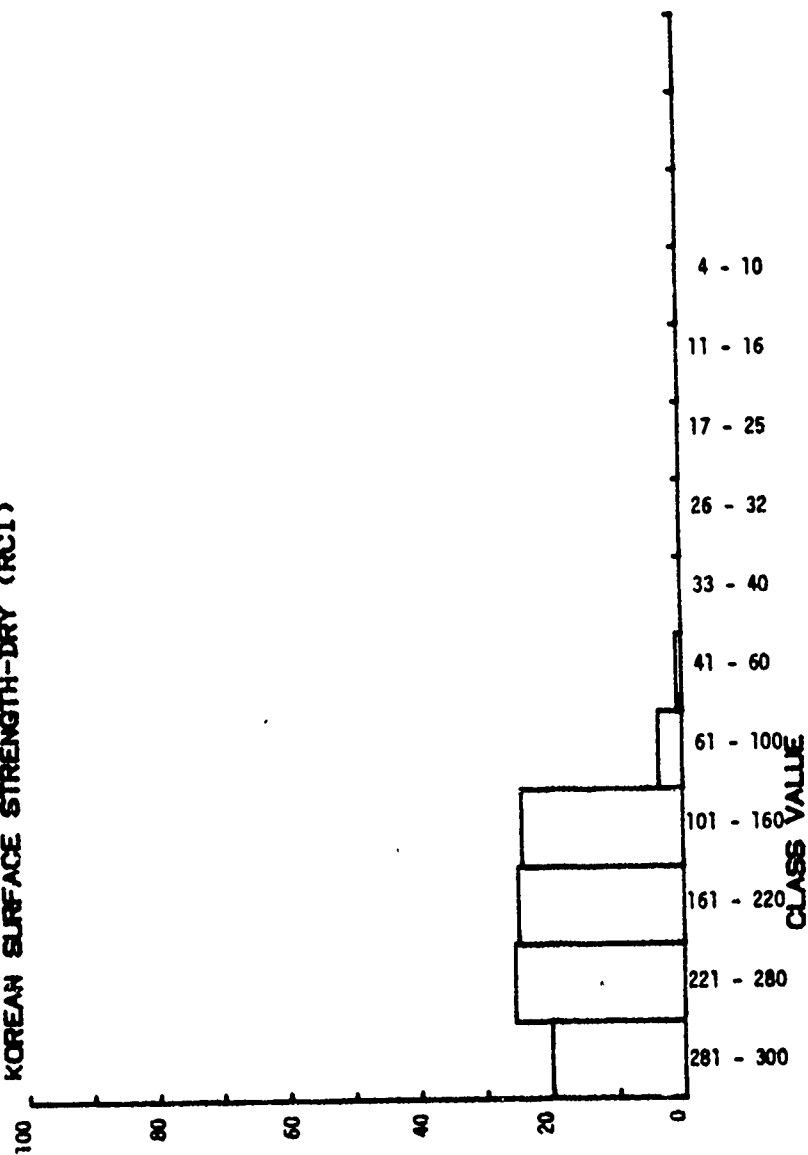


FIGURE A.2
KOREAN SURFACE STRENGTH-DRY (RCI)

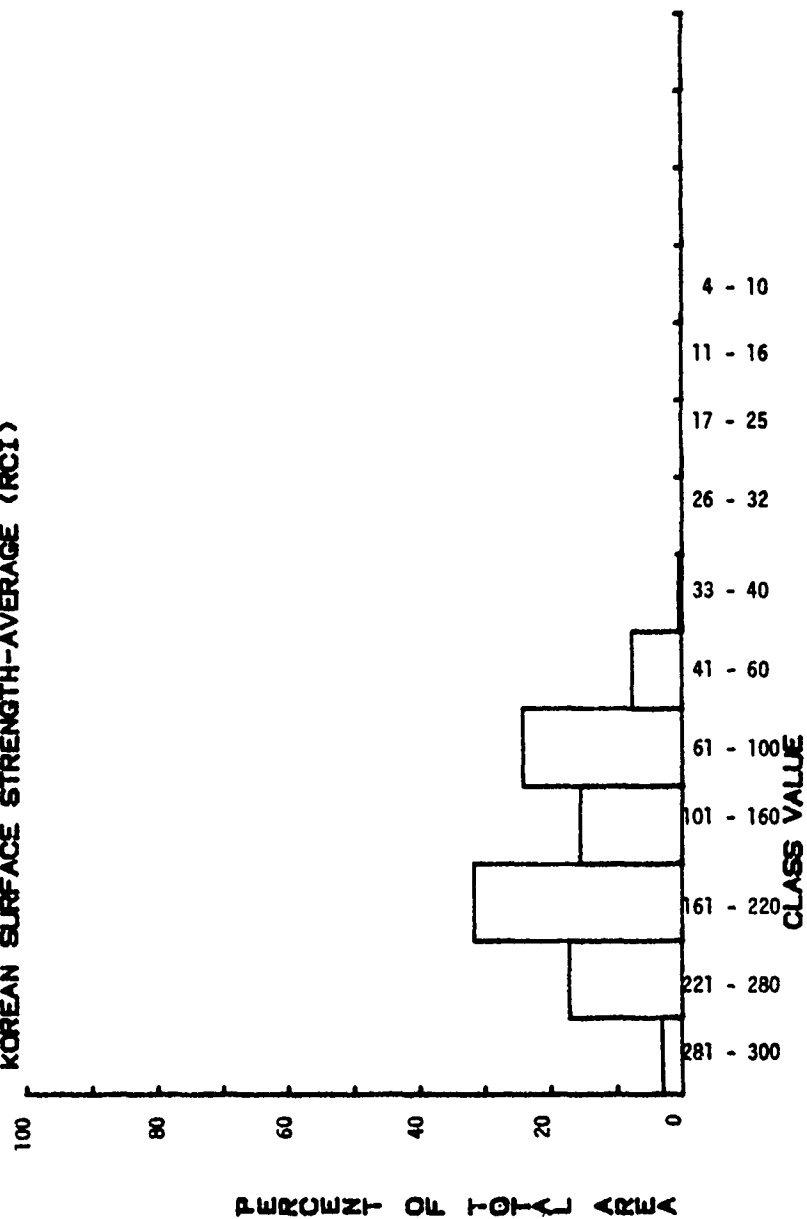


PERCENT OF TOTAL AREA

NOTE: Table values are percent of total area for which factor constrains mobility

FIGURE A.3

KOREAN SURFACE STRENGTH-AVERAGE (RCI)



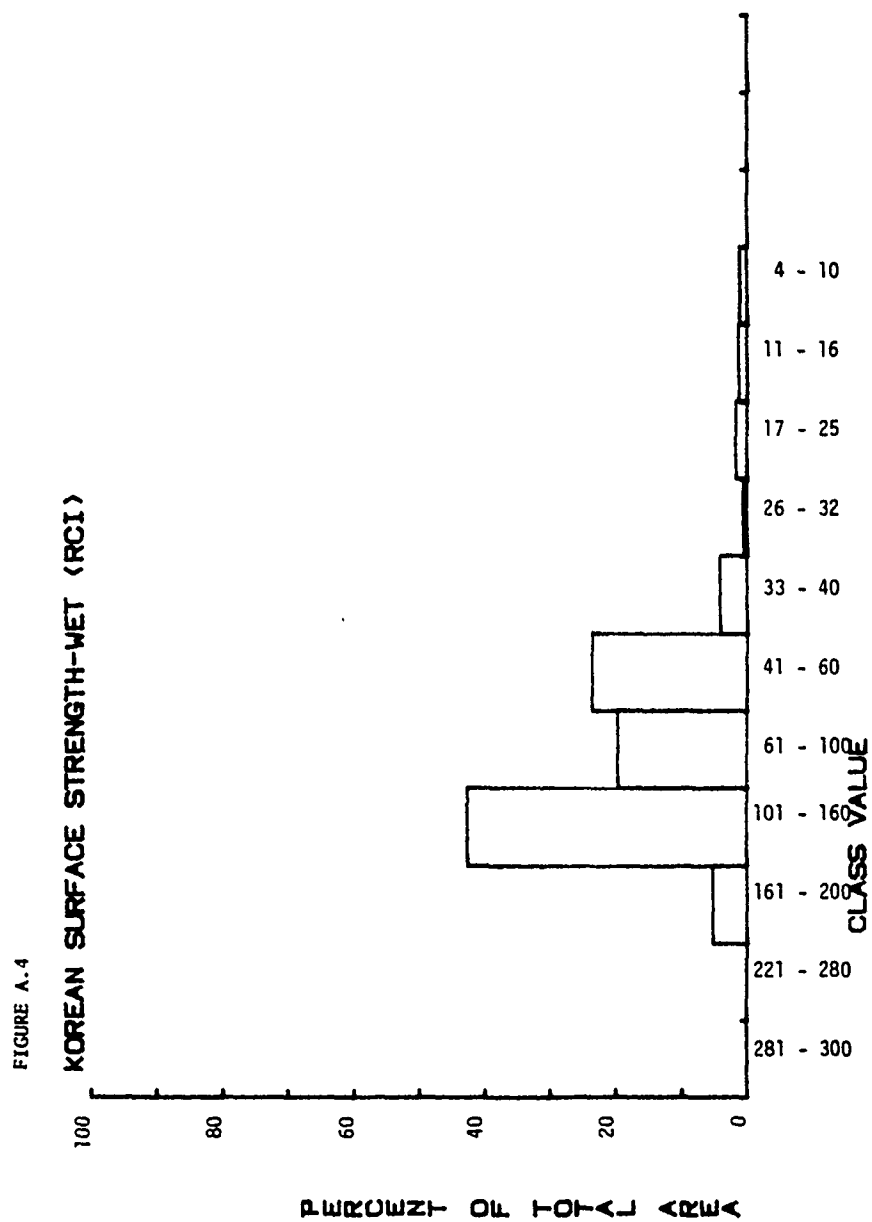


FIGURE A-5

KOREAN SURFACE STRENGTH-WETWET (RCI)

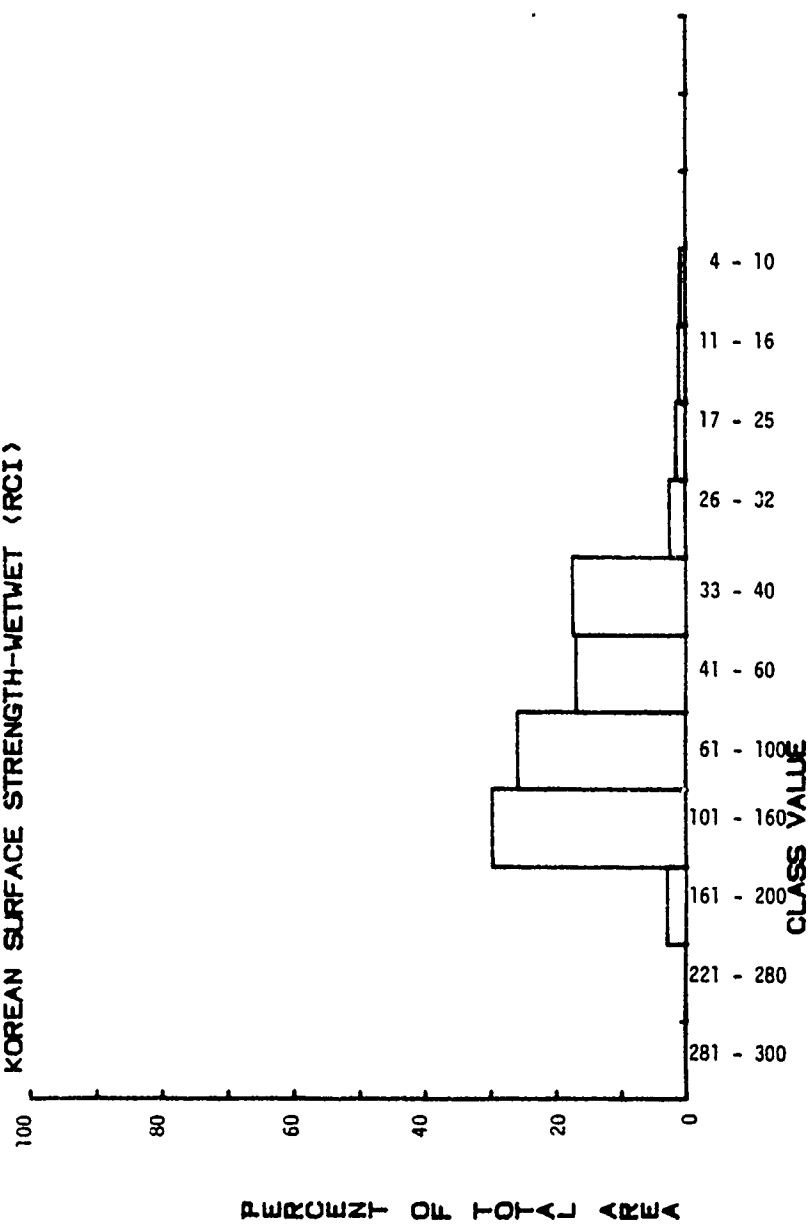
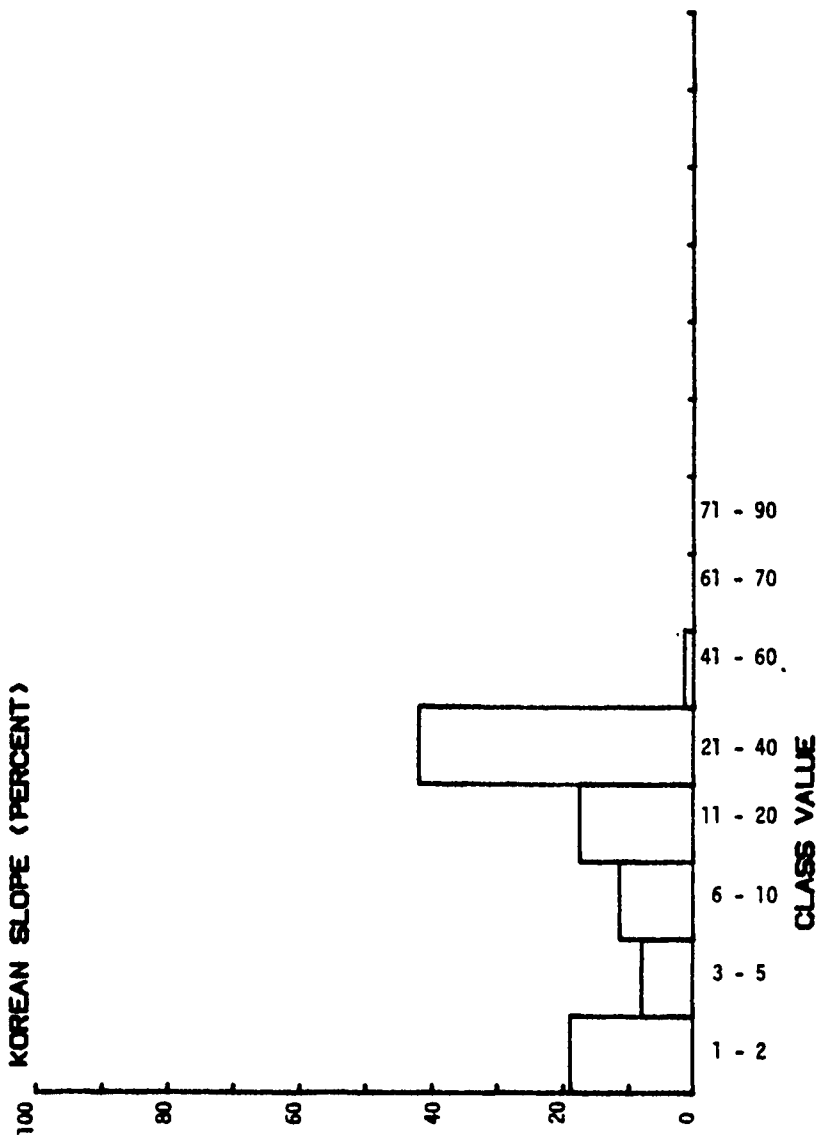


FIGURE A-6

KOREAN SLOPE (PERCENT)



PERCENT OF TOTAL AREA

FIGURE A-7

KOREAN OBSTACLE APPROACH ANGLES (DEGREES)

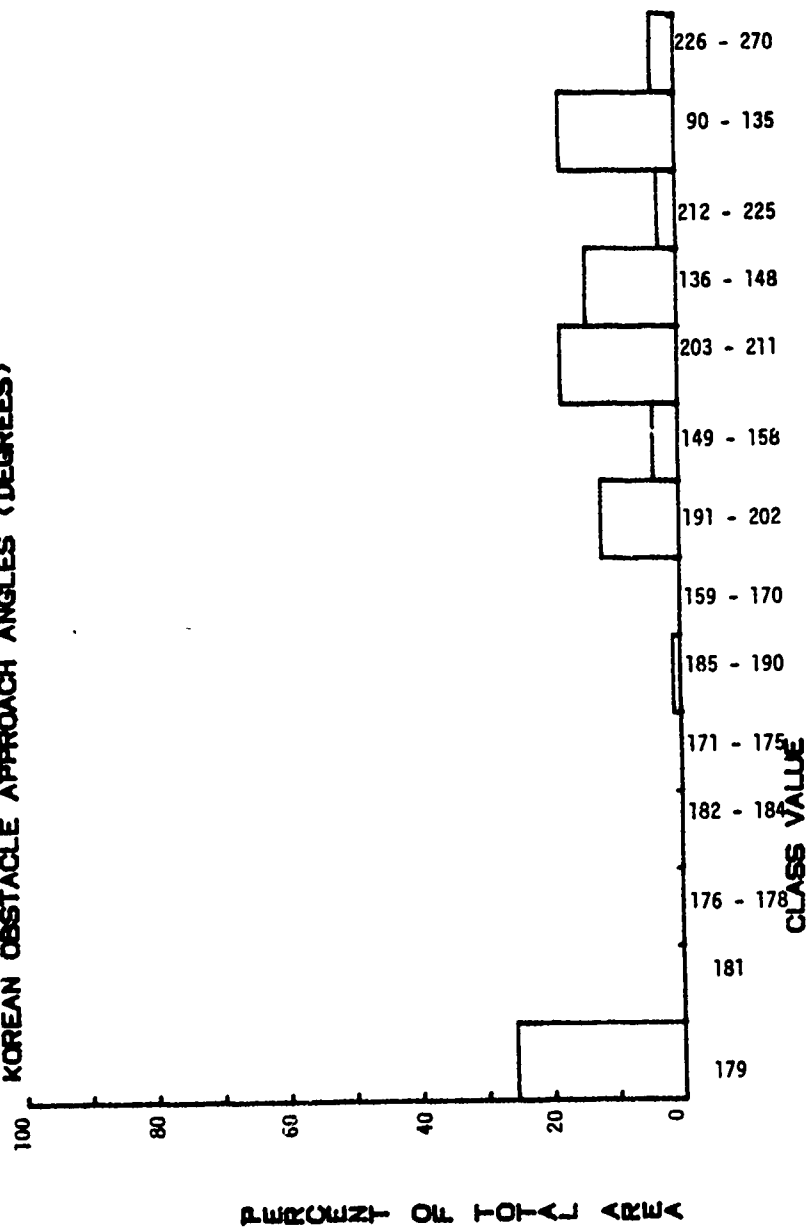
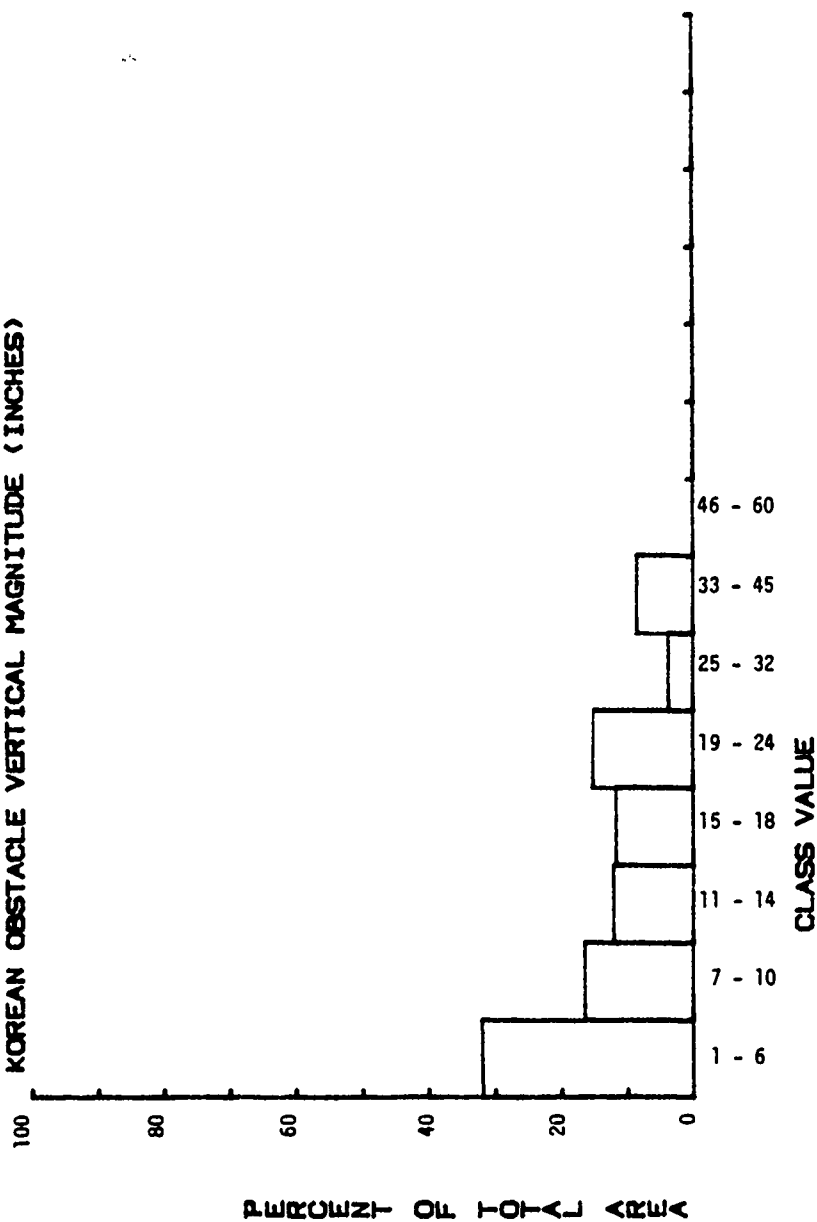


FIGURE A-8
KOREAN OBSTACLE VERTICAL MAGNITUDE (INCHES)



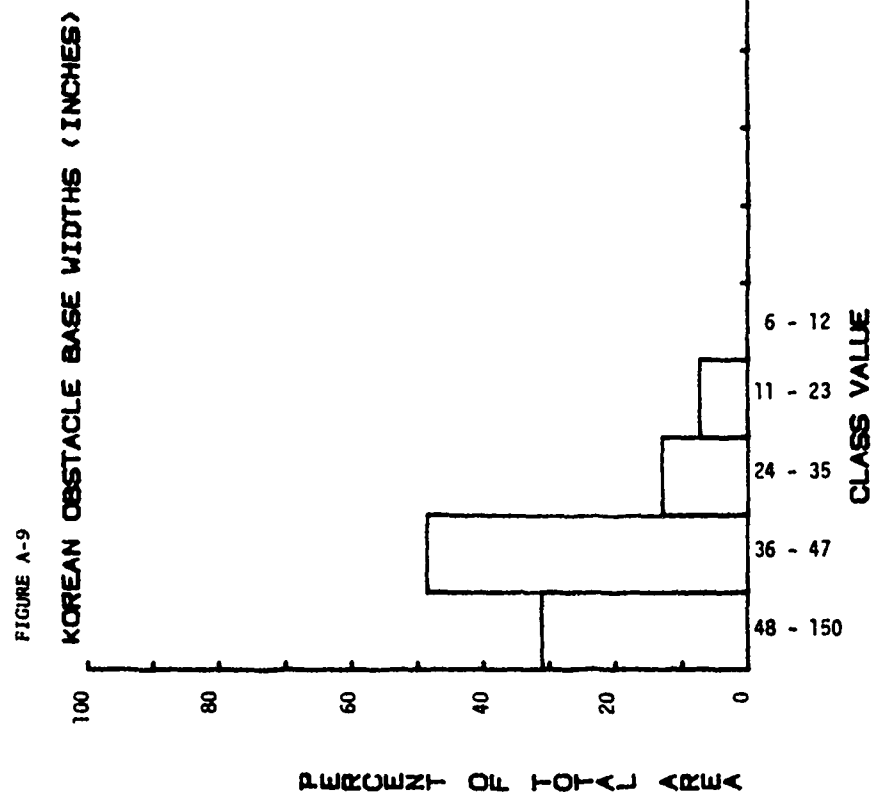


FIGURE A-10

KOREAN OBSTACLE LENGTHS (FEET)

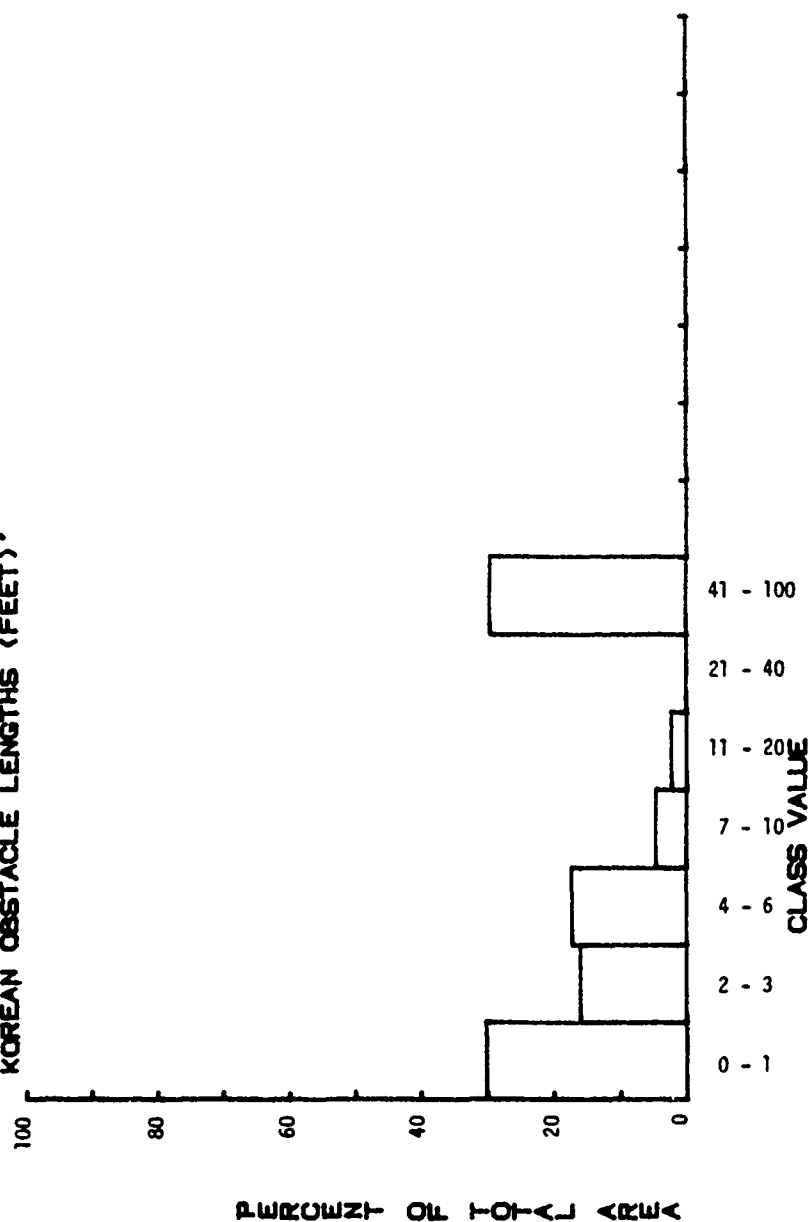
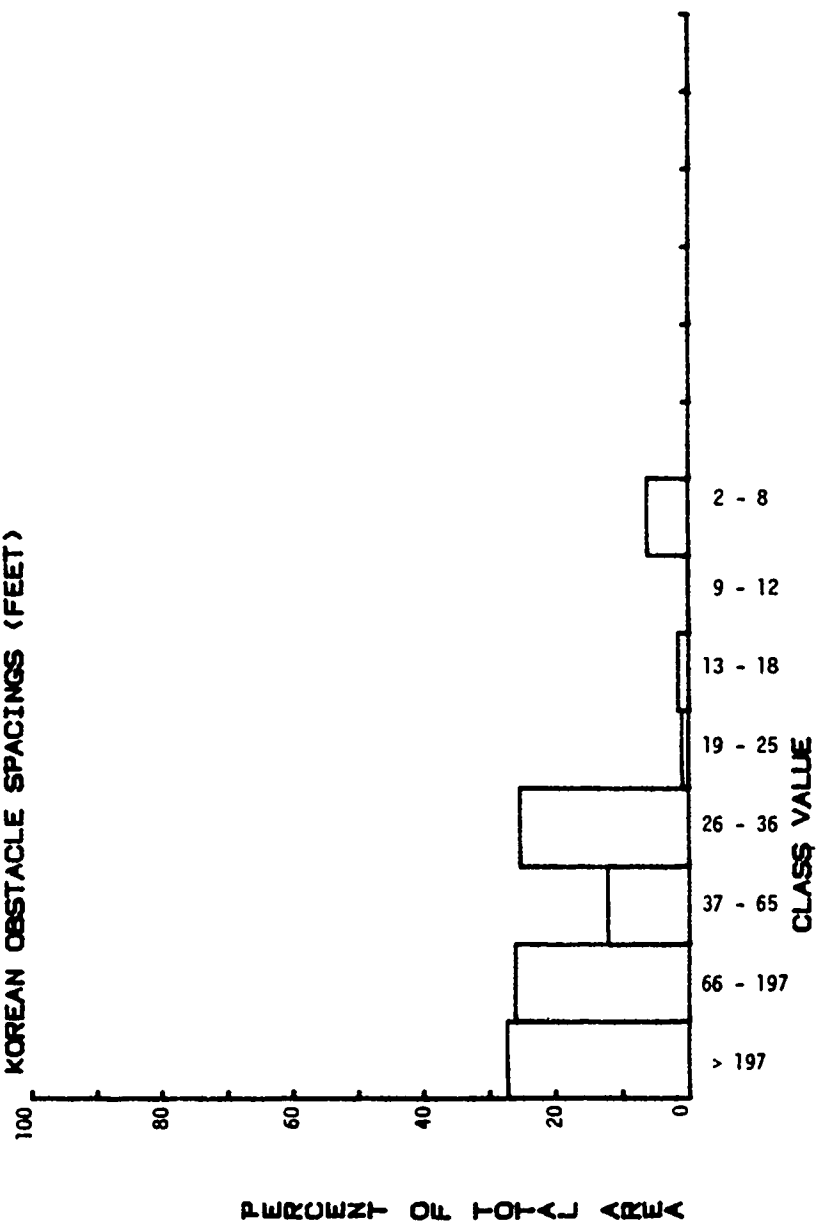


FIGURE A-11

KOREAN OBSTACLE SPACINGS (FEET)



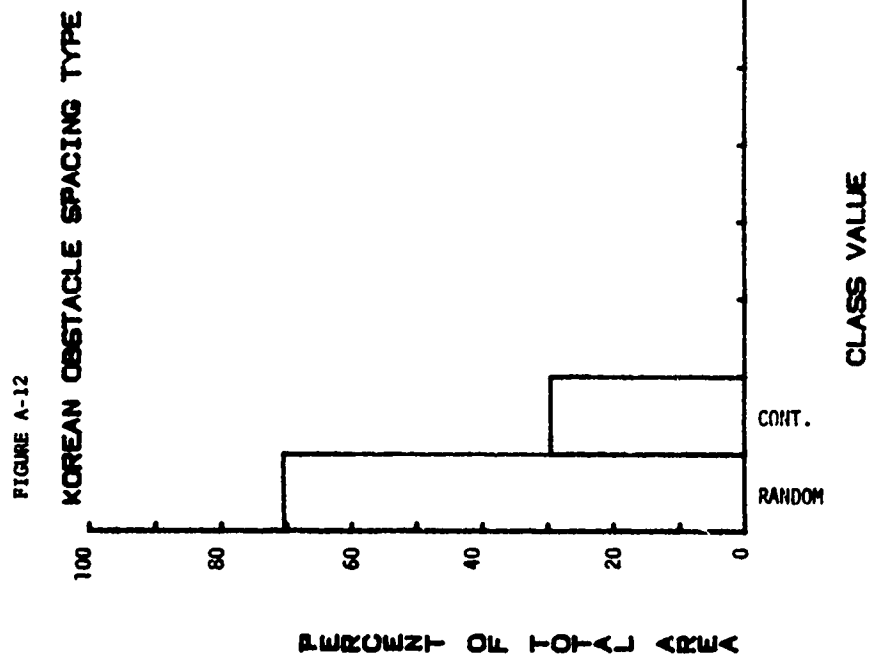


FIGURE A-13

KOREAN SURFACE ROUGHNESS (RMS X 10)

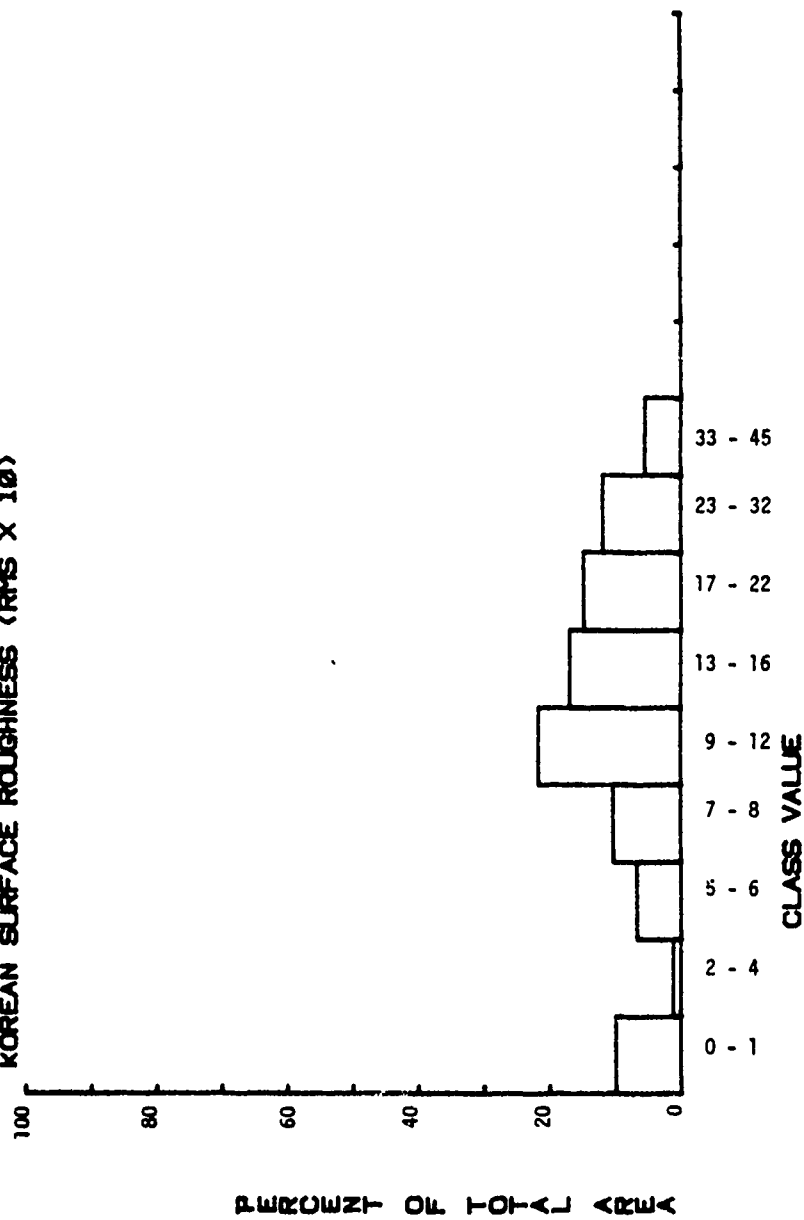


FIGURE A-14

KOREAN STEM SPACINGS > CLASS I (FEET)

(STEM DIA > 0 IN)

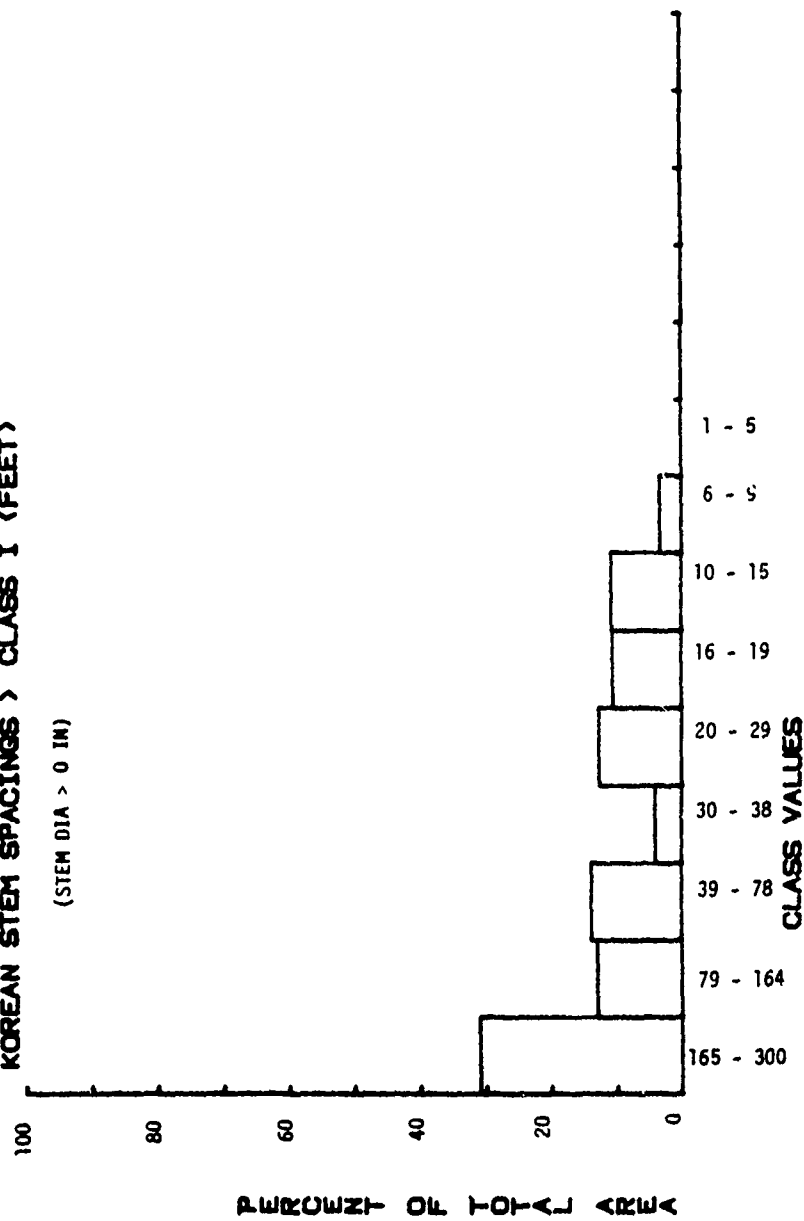


FIGURE A-15

KOREAN STEM SPACINGS > CLASS II (FEET)

(STEM DIA > 1.0 IN)

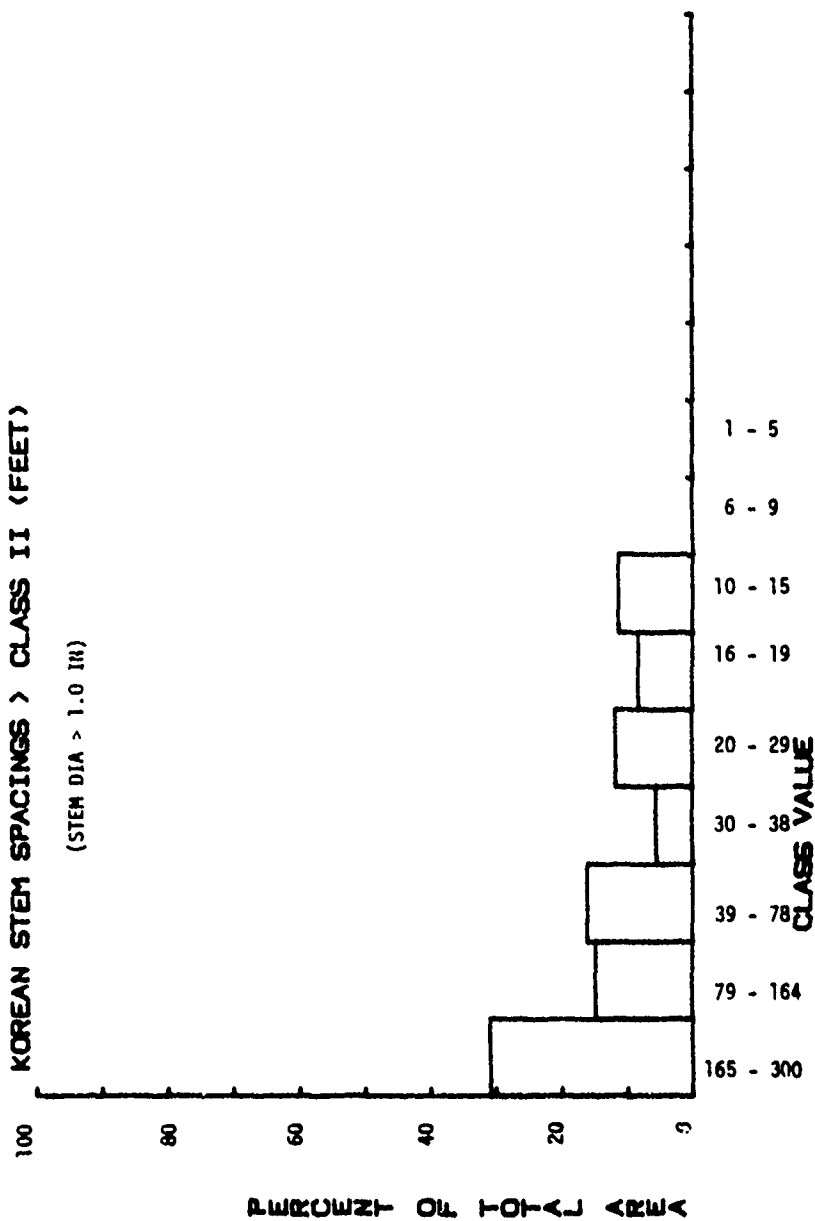


FIGURE A-16

KOREAN STEM SPACINGS > CLASS III (FEET)

(STEM DIA > 2.4 IN)

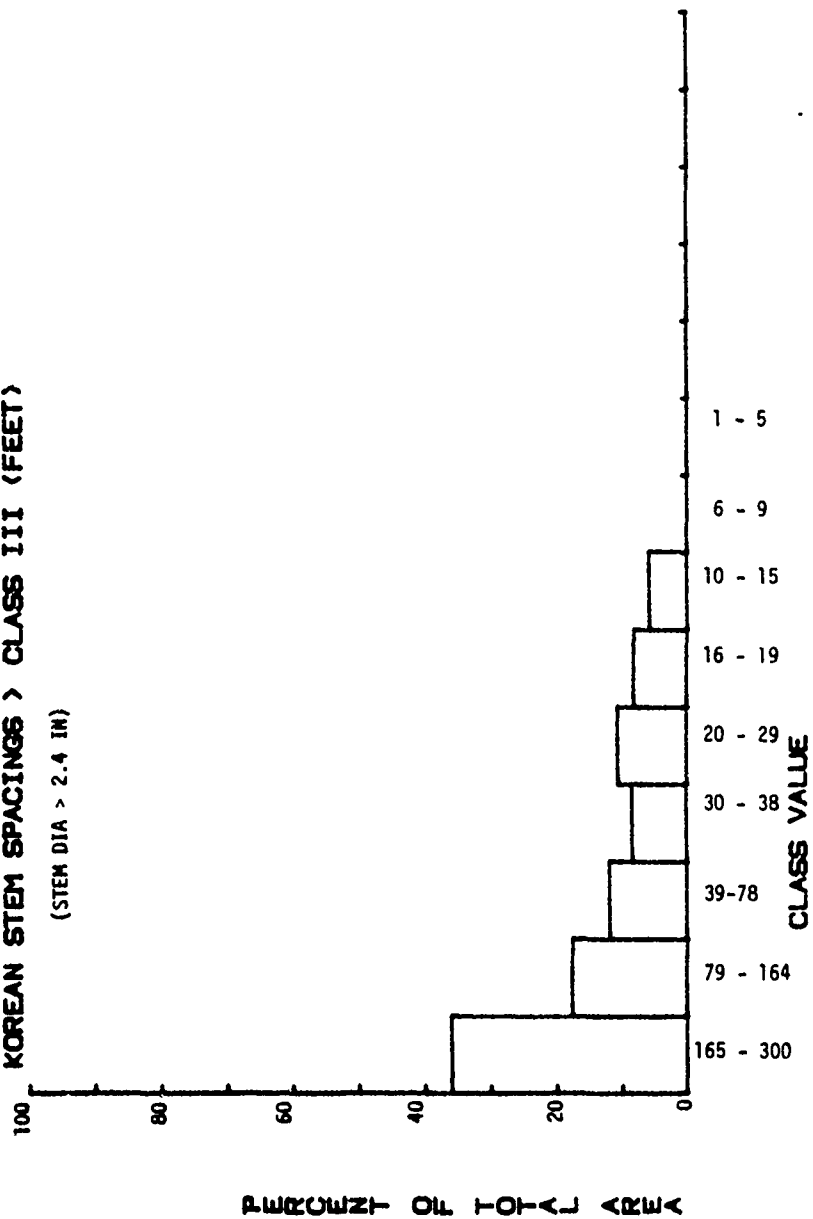


FIGURE A-17

KOREAN STEM SPACINGS > CLASS IV (FEET)

(STEM DIA > 3.9 IN)

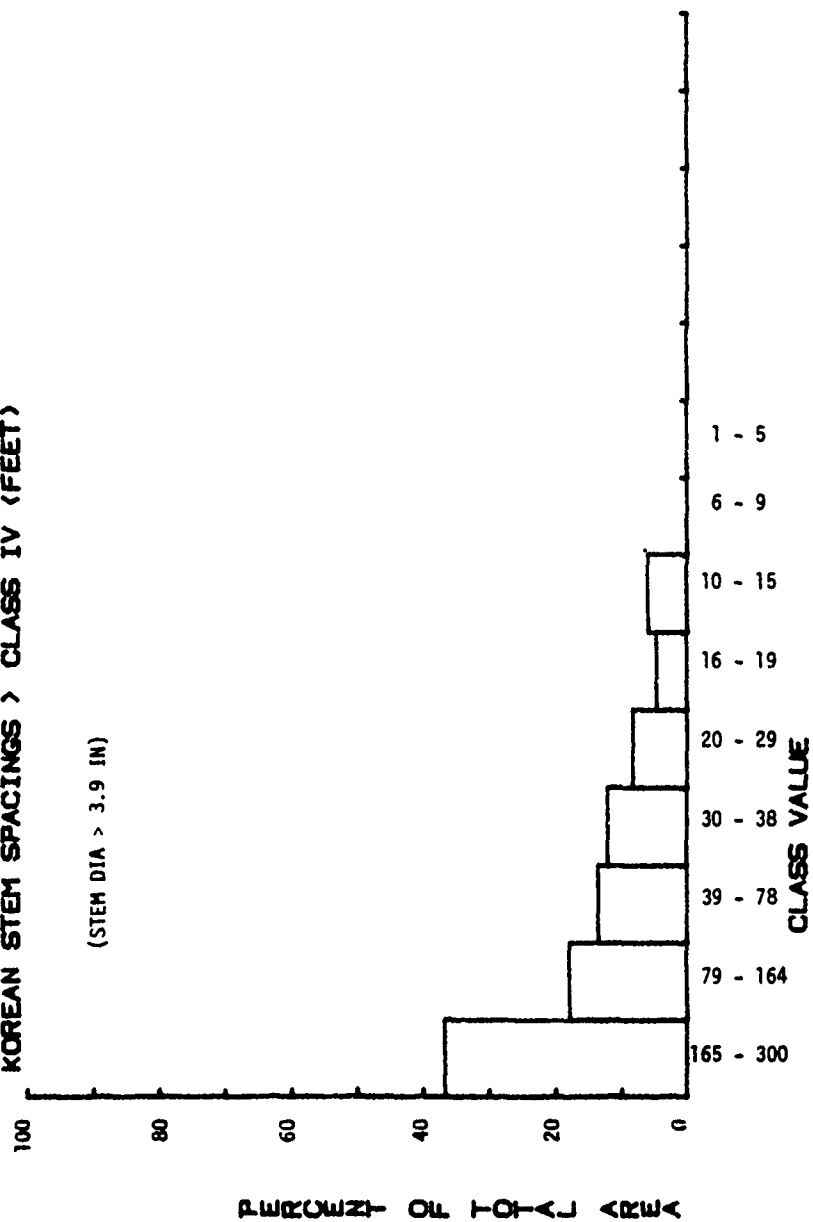
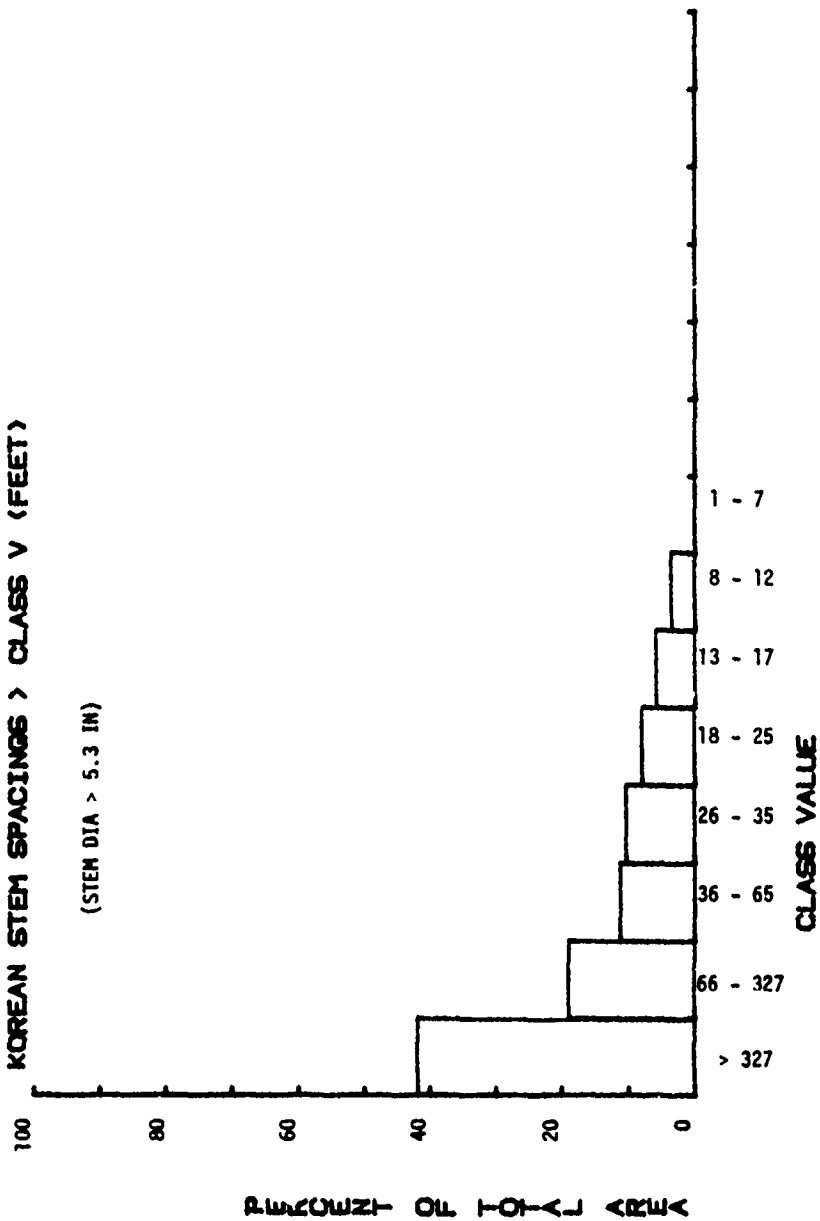


FIGURE A-18

KOREAN STEM SPACINGS > CLASS V (FEET)

(STEM DIA > 5.3 IN)



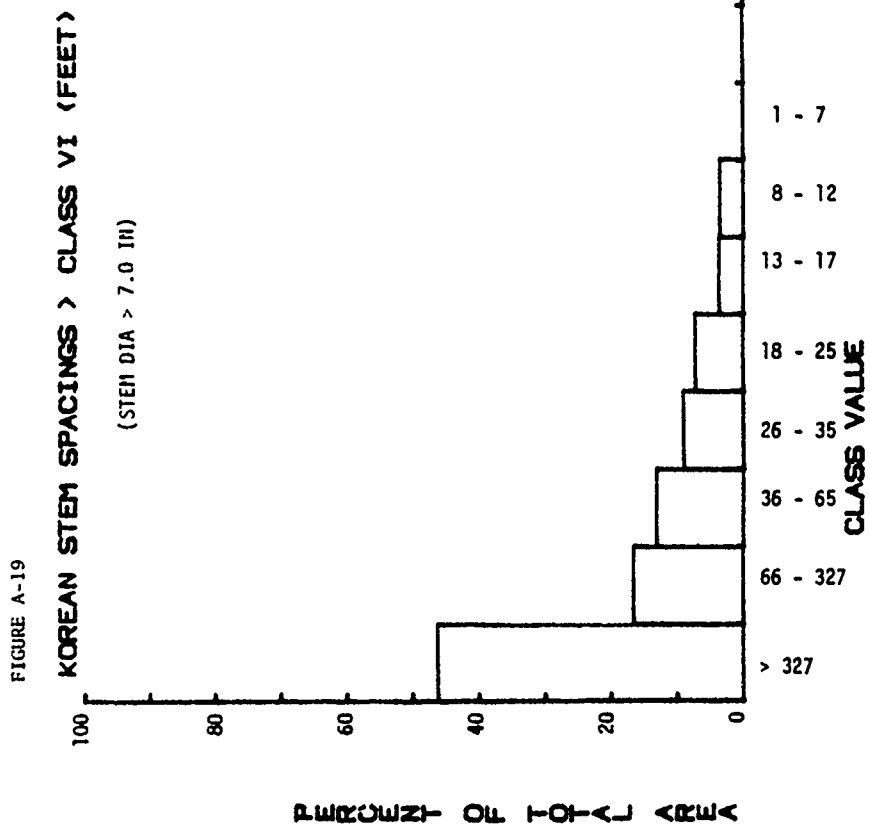


FIGURE A-20

KOREAN STEM SPACINGS > CLASS VII (FEET)

(STEM DIA > 8.7 IN)

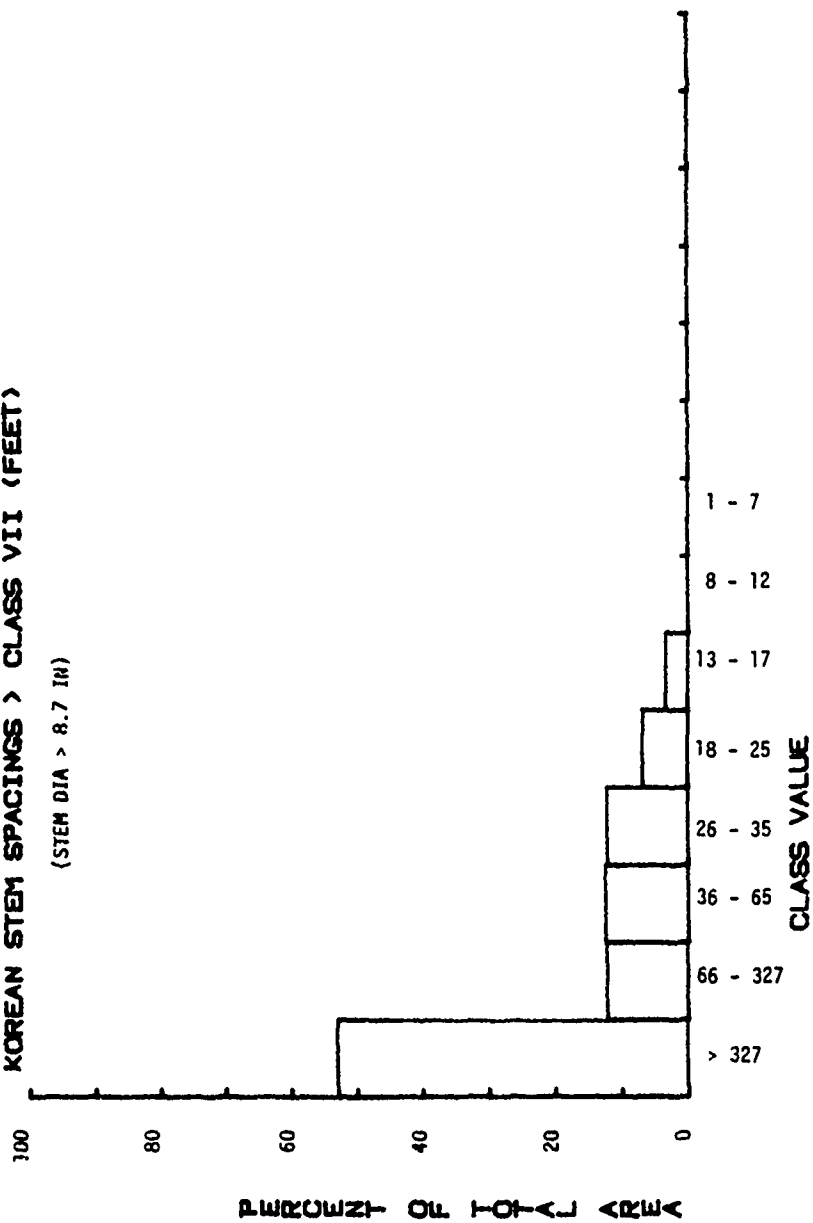


FIGURE A-21

KOREAN STEM SPACINGS > CLASS VIII (FEET)

(STEM DIA > 9.8 IN)

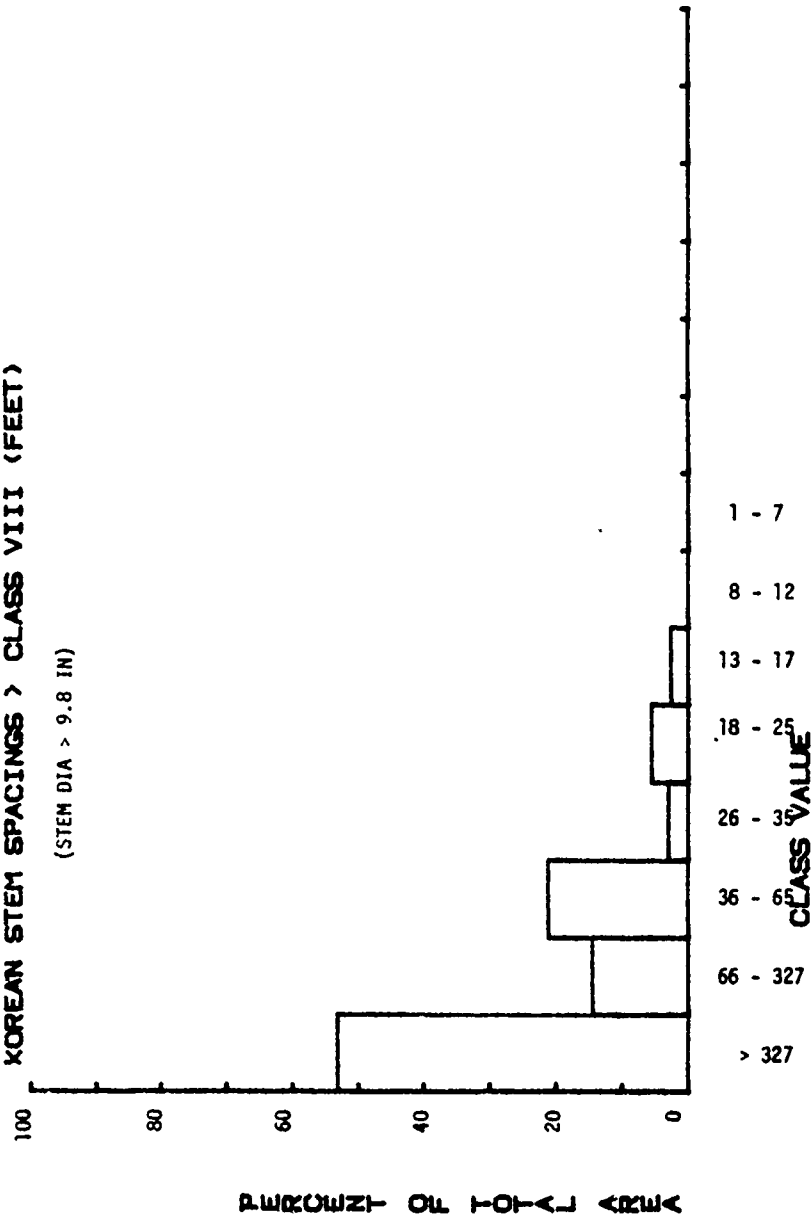


FIGURE A-22

KOREAN VISIBILITY (FEET)

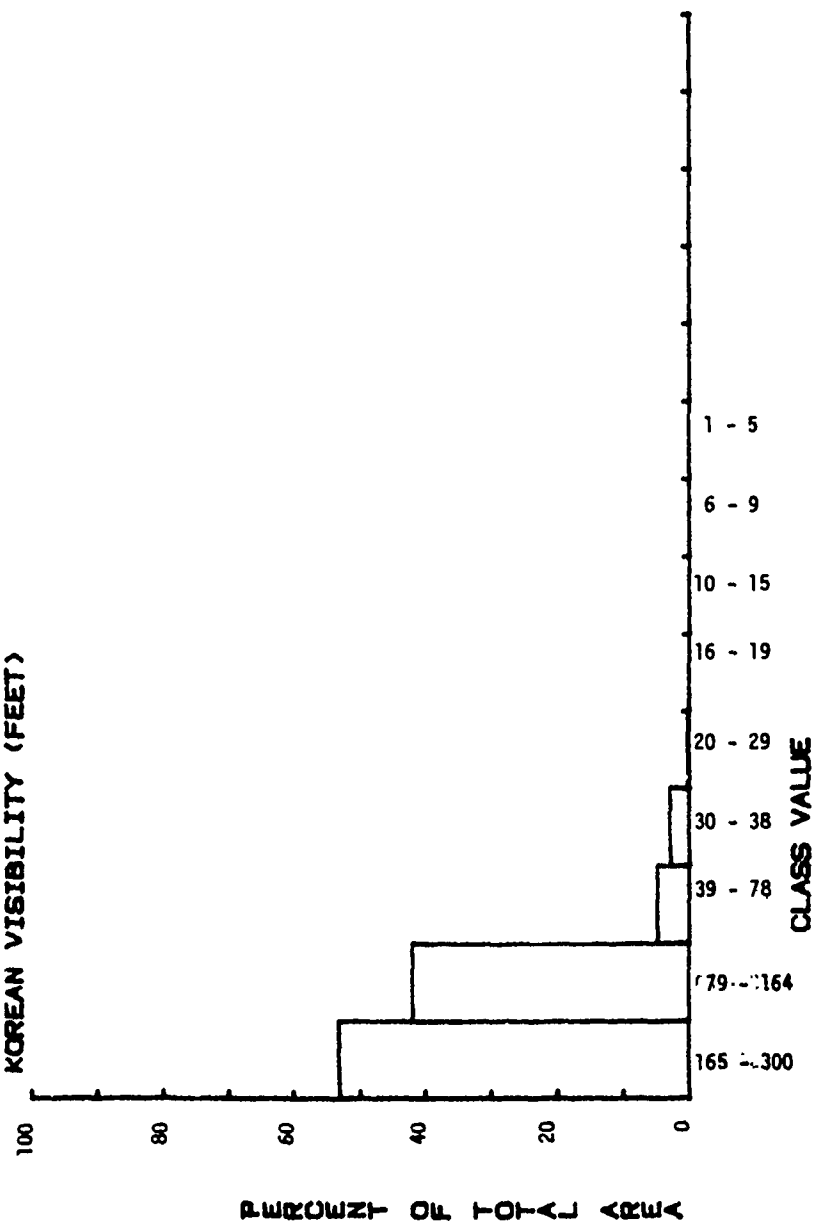
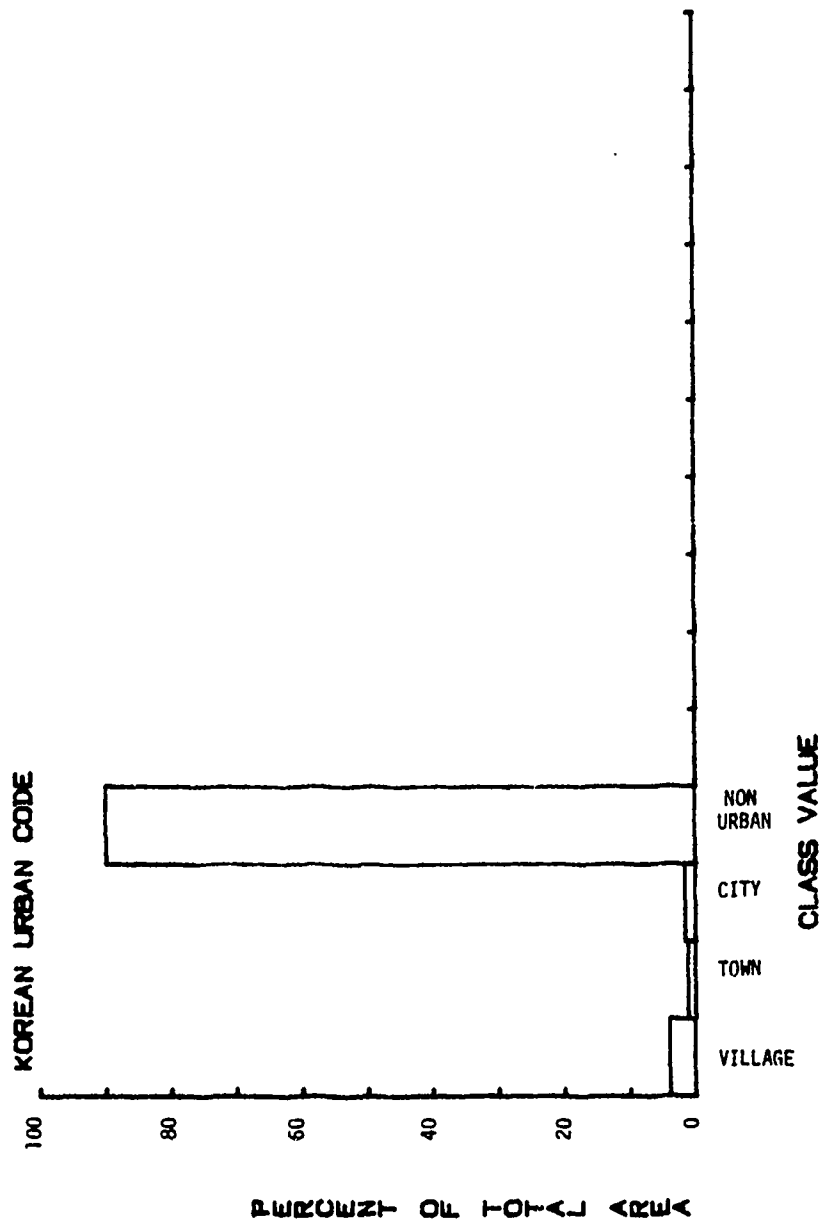


FIGURE A-23



Next page is blank

APPENDIX B

VEHICLE CONE INDEX EQUATIONS

MOBILITY INDEX FOR SELF-PROPELLED TRACKED VEHICLES IN FINE-GRAINED SOILS

Vehicle	Weight
1965 Ford Mustang	2,800 lbs
1966 Ford Mustang	2,800 lbs
1967 Ford Mustang	2,800 lbs
1968 Ford Mustang	2,800 lbs
1969 Ford Mustang	2,800 lbs
1970 Ford Mustang	2,800 lbs
1971 Ford Mustang	2,800 lbs
1972 Ford Mustang	2,800 lbs
1973 Ford Mustang	2,800 lbs
1974 Ford Mustang	2,800 lbs
1975 Ford Mustang	2,800 lbs
1976 Ford Mustang	2,800 lbs
1977 Ford Mustang	2,800 lbs
1978 Ford Mustang	2,800 lbs
1979 Ford Mustang	2,800 lbs
1980 Ford Mustang	2,800 lbs
1981 Ford Mustang	2,800 lbs
1982 Ford Mustang	2,800 lbs
1983 Ford Mustang	2,800 lbs
1984 Ford Mustang	2,800 lbs
1985 Ford Mustang	2,800 lbs
1986 Ford Mustang	2,800 lbs
1987 Ford Mustang	2,800 lbs
1988 Ford Mustang	2,800 lbs
1989 Ford Mustang	2,800 lbs
1990 Ford Mustang	2,800 lbs
1991 Ford Mustang	2,800 lbs
1992 Ford Mustang	2,800 lbs
1993 Ford Mustang	2,800 lbs
1994 Ford Mustang	2,800 lbs
1995 Ford Mustang	2,800 lbs
1996 Ford Mustang	2,800 lbs
1997 Ford Mustang	2,800 lbs
1998 Ford Mustang	2,800 lbs
1999 Ford Mustang	2,800 lbs
2000 Ford Mustang	2,800 lbs
2001 Ford Mustang	2,800 lbs
2002 Ford Mustang	2,800 lbs
2003 Ford Mustang	2,800 lbs
2004 Ford Mustang	2,800 lbs
2005 Ford Mustang	2,800 lbs
2006 Ford Mustang	2,800 lbs
2007 Ford Mustang	2,800 lbs
2008 Ford Mustang	2,800 lbs
2009 Ford Mustang	2,800 lbs
2010 Ford Mustang	2,800 lbs
2011 Ford Mustang	2,800 lbs
2012 Ford Mustang	2,800 lbs
2013 Ford Mustang	2,800 lbs
2014 Ford Mustang	2,800 lbs
2015 Ford Mustang	2,800 lbs
2016 Ford Mustang	2,800 lbs
2017 Ford Mustang	2,800 lbs
2018 Ford Mustang	2,800 lbs
2019 Ford Mustang	2,800 lbs
2020 Ford Mustang	2,800 lbs
2021 Ford Mustang	2,800 lbs
2022 Ford Mustang	2,800 lbs
2023 Ford Mustang	2,800 lbs
2024 Ford Mustang	2,800 lbs
2025 Ford Mustang	2,800 lbs
2026 Ford Mustang	2,800 lbs
2027 Ford Mustang	2,800 lbs
2028 Ford Mustang	2,800 lbs
2029 Ford Mustang	2,800 lbs
2030 Ford Mustang	2,800 lbs

Track Description

$$\text{Mobility Index} = \left[\frac{(1) \times (2)}{(3) \times (4)} + (5) - (6) \right] \times (7) \times (8)$$

- | | | | | | | | | |
|-----|-------------------------|---|--|---|-------|---|-------|-----|
| (1) | Contact Pressure Factor | = | $\frac{\text{Gross weight, lb}}{\text{Area of tracks in contact with ground, sq in.}}$ | = | _____ | = | _____ | (1) |
| (2) | Weight Factor | : | $<50,000 \text{ lb} = 1.0$
$50,000 \text{ to } 69,999 \text{ lb} = 1.2$
$70,000 \text{ to } 99,999 \text{ lb} = 1.4$
$100,000 \text{ lb or } > = 1.3$ | = | _____ | | | (2) |
| (3) | Track Factor | = | $\frac{\text{Track width, in.}}{100}$ | = | _____ | = | _____ | (3) |
| (4) | Grouser Factor | : | $<1.5 \text{ in. high} = 1.0$
$>1.5 \text{ in. high} = 1.1$ | = | _____ | | | (4) |
| (5) | Bogie Factor | = | $\frac{\text{Gross wt} + 10}{\text{Total no. bogies in contact with ground} \times \text{area of 1 track shoe}}$ | = | _____ | = | _____ | (5) |
| (6) | Clearance Factor | = | $\frac{\text{Clearance, in.}}{10}$ | = | _____ | = | _____ | (6) |
| (7) | Engine Factor | : | $>10 \text{ hp/ton} = 1.00$
$<10 \text{ hp/ton} = 1.05$ | = | _____ | | | (7) |
| (8) | Transmission Factor | = | $\text{Hydraulic} = 1.00$
$\text{Mechanical} = 1.05$ | = | _____ | | | (8) |

$$\text{Mobility Index (MI)} = \left[\frac{x}{x} + \dots \right] \times x = \dots$$

$$VC_{I50} = 19.27 + .43 MI - \left[\frac{125.79}{MI + 7.08} \right]$$

VCI₅₀ = 19.27 +

$$VCI_1 = 7.0 + .2MI - \left[\frac{39.2}{MI + 5.6} \right]$$

7.0 + []

Next page is blank

FIRE CONTROL SYSTEM PERFORMANCE DEGRADATION
WHEN A TANK GUN ENGAGES A MANEUVERING THREAT

JOHN J. MCCARTHY

HAROLD H. BURKE

US Army Materiel Systems Analysis Activity
Aberdeen Proving Ground, MD 21005, U.S.A.

ABSTRACT. Current specifications to assess the performance of tank gun fire control systems consider a non-maneuvering vehicle as the threat. With the introduction of highly mobile and agile vehicles to the battlefield environment this type of vehicle movement is a subset of the potential maneuvering that the threat is capable of executing.

Results of an investigation to determine the degradation in performance of fire control systems engaging maneuvering threats are presented. Modeling of both the movement characteristics of the threat and the possible alternate types of fire control systems (i.e., manual, disturbed reticle, stabilized sight-director) are discussed.

Existing tactical vehicles are shown to exhibit mobility and agility characteristics which when combined with projectile time of flight cause the performance of predictive gun fire control systems to degrade in their ability to provide accurate firepower. The identification of the sources of projectile miss distance are identified as the tracking, estimation and prediction processes and quantified results are presented.

The potential payoff in performance realized when non-linear prediction is incorporated into the design of predictive fire control systems is discussed. The design challenge to develop an effective suboptimal multi-variable fire control system is acknowledged and the benefits derived from the application of such a design methodology are compared to the performance limitations of existing gun fire control system technology.

1. INTRODUCTION

This paper discusses the nature of land vehicle mobility and agility and explores the ability of different types of lead predictive fire control systems to effectively engage such vehicle maneuvering. Existing performance specifications do not satisfactorily describe in quantitative terms the level of maneuverability to be expected. Rather, the present specifications define performance requirements for fixed vehicle speed and heading movement. This type of requirement has motivated the development of fire control system designs that are significantly degraded for a maneuvering threat environment.

In order to develop a realistic set of performance specifications that will stand firm when compared to real world vehicle maneuvering, considerable care must be exercised to establish a maneuvering vehicle criteria. The quantitative description should be such that neither undue constraints are imposed on fire control system performance nor should they be so lax as to exclude realistic threat maneuvering capabilities.

The present discussion is organized to present the following material: (a) a quantitative model describing the nature of land vehicle maneuvering, (b) different possibilities of tank gun fire control system design, (c) sources and magnitudes of pointing errors when maneuvering threats are engaged, (d) the utilization of firing doctrine to improve performance and (e) an alternate methodology for the development of fire control systems that offers significant performance improvement. The information presented will contribute to an increased understanding of land vehicle mobility and agility, describe the degrading influence of maneuvering threats, and suggest an approach that will provide an improvement in fire control effectiveness against maneuvering threats.

2. MODELING THE MANEUVERING THREAT

2.1 Maneuvering Vehicle Modeling Methodology

The critical motion parameters of maneuvering vehicle paths that degrade the performance of predictive fire control systems have been identified in an AMSAA report [1] as cyclic oscillations exhibiting frequencies that are within the maneuvering capabilities of tactical land vehicles.

Tracking error, defined as the difference between target and reticle position, does not in itself cause the performance degradation. The inability of the fire control system to determine the motion derivatives of the line-of-sight to the target, and predict the future position of the target are

the two main factors that cause predictive fire control systems to degrade when engaging maneuvering targets.

Improvements in the performance of fire control systems operating against maneuvering targets can be realized by upgrading the ability to obtain a better estimate of the threat's line-of-sight movement and predict its future position as shown in Fig. 2.1. A quantitative description of the maneuvering threat is needed to evaluate the extent of the performance degradation. To develop this description it is necessary to consider the mobility and agility characteristics of threat vehicle movements. A thorough description of anticipated maneuvering seems to defy identification because threat maneuvers constitute a large set of possibilities even when constrained by tactical doctrine and vehicle capabilities. An analytic approach to describing maneuvers would view each maneuver as being composed of elements from an idealized group of movements. An empirical approach would view the maneuvers as having actually occurred during limited tests of different types of maneuvering vehicles. Neither of these approaches provide a complete maneuver description, but a combination of these two approaches offers some advantages and is the rationale adopted. The analytic approach will partially overcome the incompleteness of the empirical data base while the empirical data will offset the mathematical idealizations of the analytic methodology.

2.2 Empirical Approach

When using empirical data to demonstrate the performance of a gun fire control system, baseline performance can be determined with no concerns arising from idealization of the maneuvers. Since the number of maneuvers will be rather small, they neither provide sufficient information about the robustness of a fire control design methodology nor the pathology when the fire control system begins to degrade. When demonstrating the performance of a fire control system against experimental data, caution must be exercised to assure that the empirical data is properly inputted to the fire control system. Matching of the data rates and noise levels often requires some preprocessing of experimental data to prepare it for use in simulation studies.

A non-maneuvering vehicle path and three maneuvering profiles shown in Fig. 2.2 characterize the motion characteristics of three types of maneuvering vehicles; an M60, Scout and Twister. The position location data of the maneuvering vehicles is obtained at a 10/sec rate which is required for the accurate determination of the rates and accelerations needed for the fire control system performance studies.

2.3 Analytic Approach

As a supplement to the empirical approach the analytic approach is used to investigate sensitivity effects for a larger group of movements. Simulating new or pathological maneuvers requires that the analytic capability superimpose maneuvers arising from random disturbances and intentional, voluntary vehicle driver commands.

2.3.1 Random Disturbances. The random disturbances may be represented in terms of time histories or power spectral densities. The time history approach is based on the development of a mathematical model of vehicle movement influenced by terrain effects and arbitrary driving habits of individual drivers. It is assumed that for no random effects caused by terrain irregularities or drive input, the vehicle would follow a straight line-constant speed path. Maneuvers are viewed as perturbations on this straight line-constant speed path. Apparent acceleration $a(t)$, accounts for the vehicle's deviation from a straight line path. Maneuver capability is expressed by three quantities: the variance, or magnitude of $a(t)$, the cyclic maneuver frequency and the time constant of the maneuver.

The apparent vehicle acceleration is correlated in time; if the target is accelerating at time t , it is likely to be accelerating at $t+\tau$ for sufficiently small τ . A representative model of the correlation function, $\phi(\tau)$, associated with the apparent acceleration is

$$\begin{aligned}(\phi)\tau &= E [a(t)a(t+\tau)] \\ &= \sigma^2 e^{-\alpha \tau} \cos \omega_1 |\tau|\end{aligned}$$

where α = reciprocal of maneuver (acceleration time constant)
 ω_1 = cyclic frequency of maneuver acceleration.
 σ^2 = variance of maneuver acceleration.

The auto correlation function for a periodic random acceleration is shown in Fig. 2.3.

The vehicle can accelerate at a maximum rate A_{\max} ($-A_{\max}$) and will do each with a probability density of

$$\sigma^2 = \frac{A_{\max}^2}{3} [1 + 4 P_{\max} - P_0]$$

The assumed acceleration probability density is shown in Fig. 2.4.

A block diagram of a periodic random acceleration generator is shown in Fig. 2.5.

The power spectral density representation is an alternate way to express the random acceleration in the frequency domain. The form of the power spectral density in terms of the parameters used in the time domain approach is

$$\phi(\omega) = 2\sigma^2\alpha \frac{\omega^2 + (\alpha^2 + \omega_1^2)}{\omega^4 + 2(\alpha^2 - \omega_1^2)\omega^2 + (\alpha^2 + \omega_1^2)^2}$$

The shape of the power spectral density curve depends on the quantity $3\omega_1^2 - \alpha^2$ and is shown in Fig. 2.6. The peaking of the acceleration power spectral density as shown in case b of Fig. 2.6 is important because this set of maneuvering vehicle parameters yields the cyclic oscillations that have been shown in Reference [1] to be responsible for the degradation of fire control system performance.

Representative motion is synthesized by reconstructing the time history from the power spectral density. This requires a random selection of phases for the various frequency components. The motion that is created has stationary statistical properties and thus approximates the time history case with the imposed limitations of a constant σ^2 and steady state operation.

2.3.2 Intentional, Voluntary Vehicle Driver Commands. Motion of land vehicles over terrain is a complicated subject in itself and will not be investigated in this study. It is recognized however, that an interaction between vehicle horsepower, weight, suspension, and locomotion concepts do combine with terrain over which it is moving to provide different levels of mobility with respect to a fixed reference frame. Therefore, different vehicle designs will have different mobility levels defined in terms of motion and derivatives of motion.

Agility is closely related to mobility and yet it is a slightly different description of vehicle motion. Where mobility describes the movement of a vehicle from one location to another location in a given period of time, agility describes the vehicle's ability to alter its mean path during that time period. An example of these two parts of vehicle motion would be to observe that a tracked vehicle travels from A to B in 100 seconds (mobility) while it executes several deceleration, acceleration and oscillatory movements (agility).

2.3.3 Parameters of Typical Mobile/Agile Vehicles and Projectiles. The speed range of vehicles that will be considered is $0 < A < 3$ meters/ S^2 , and the linear deceleration is: $0 < \bar{A} < 10$ meters/ S^2 . The radial acceleration is $0 < \bar{A} < 10$ meters/ S^2 . The time of flight of the projectile is a function of range between the fire control system and the evasive vehicle and is $0.5 \leq t \leq 4$ sec. The cyclic occurrence of the

apparent motion of the evasive vehicle is $0 \leq f \leq 1/4$ Hz.

2.4 Model of Maneuvering Vehicle Motion

The path traversed by a maneuvering vehicle proceeding from point A to point B can be thought of as being generated by moving on circular arcs connected by essentially straight line segments as shown in Fig. 2.7. The radii of these circles vary from $10 < R < \infty$ meters. The speed of the evasive vehicle is < 15 meters/S as it moves thru the circular arcs. The number of circular arcs moved on while moving from point A to point B, the total angle of the circular arc, the radius of each and the speed variation over the interconnecting straight segments between the circular arcs determines the mobility/agility of the evasive vehicle. Orientation of the engaging gun fire control system will establish a reference location from which the apparent cyclic motion of the evasive vehicle can be observed. Fig. 2.7 describes the relationships that exist for head-on maneuvering of a threat vehicle. The apparent velocities and accelerations are considerably different than for a simple sinusoidal motion model. Fig. 2.8 shows the motion parameters for a typical head-on engagement and Fig. 2.9 describes the apparent motion resulting from a serpentine maneuver.

The random and intentional accelerations are summed to obtain a resultant forcing as shown in Fig. 2.10. The apparent motion from the combined models may be used in the place of empirical data to simulate input to a fire control system. The effects of vehicle parameter variations are readily observed without resorting to additional field testing. A maneuvering threat path generator is under development that will simulate both the random and intentional movement characteristics of vehicles.

3. FIRE CONTROL SYSTEM PROCESSES

The operation of all predictive fire control systems may be thought of as occurring in three processes: tracking, estimation, and prediction. Fig. 3.1 describes the relationship of the processes in a fire control system.

Tracking is usually accomplished manually and is concerned with the alignment of the sight reticle with the target. The gunner is involved directly at this stage and accuracy of tracking will be a characterization of the ability of any given gunner to perform the task. Test data obtained from experimental investigations can be used to determine tracking error means, standard deviations, and correlation time constants useful for building models of the tracking process.

Estimation of target present position is the process of

filtering the tracking data the fire control system uses as input into the prediction process. Dependent upon the specific fire control mechanization, i.e., disturbed reticle sight or stabilized sight and the feedback loops involved, the identification of the dominant estimation time constant is established. The quality of the tracking error will influence the performance of the estimation process.

Prediction of target future position to obtain intercept between projectile and target is dependent upon an estimate of the present motion of the target and time of flight of the projectile to the target. The output of the estimator is usually not a complete description of the present motion of the threat. Therefore, in general, the predictor does not have the necessary information to theoretically calculate the threat's future position. If restrictions are placed on the allowable threat motions, then the predictor's ability to determine its future position is improved. Oversimplification of allowable threat motions has placed unrealistically simplified requirements on the operation of the estimation and prediction processes. Realistic threat motions are determined by the mobility capabilities of tactical vehicles. In the past, the majority of threats that have been studied have been nonaccelerating. The requirements of an estimator and a predictor for this type of motion are to combine the apparent threat velocity estimate and projectile time of flight. The required lead is constant and can be realized after some settling time. The existence of accelerating threats requires the estimator and predictor to develop constantly changing lead angles.

4. TYPES OF FIRE CONTROL SYSTEMS

4.1 Block Diagrams

The three basic types of fire control configurations in existence are manual, disturbed reticle and stabilized sight-director systems. They are identified in terms of how each of the three fire control processes are mechanized. All existing operational systems utilize the human operator to monitor the difference between the observed target and the reticle and null the error. The degree of participation of the human in each of the three types of fire control systems is considerably different. Concern about the stability of the closed loop man-machine system is an important consideration in determining performance and is one of the primary distinguishing features that separates the potential effectiveness of the three types of fire control systems.

Fig. 4.1 (a,b,c) shows a simplified block diagram of each of the three types of fire control systems. Areas where the tracking, estimation and prediction processes occur are

identified for each system. In the manual system all three processes are performed by the man and the machine serves only to orient the gun line in accordance with the information provided by man. The tracking is performed by the man in the disturbed reticle and stabilized sight-director systems, however it is accomplished differently. The estimation and prediction processes are also mechanized differently in these two types of fire control systems as seen in Fig. 4.1 (b&c). One of the important inherent advantages of a stabilized sight-director system compared to a disturbed reticle system is the decoupling of the tracking process from the estimation and prediction processes as shown in Fig. 4.2.

The turret and gun position serve as the reference from which the reticle is disturbed in the disturbed reticle system. Fig. 4.1b shows the involvement of the human gunner in the turret loop for the disturbed reticle system and his absence in the turret loop for the stabilized sight-director system is shown in Fig. 4.1c. The tracking process is therefore more isolated from the estimation and prediction processes in the stabilized sight-director system. The performance of the estimator is dependent upon the performance of tracking and therefore the improved tracking will enhance the ability of the fire control system to develop velocity and acceleration estimates required for predictors that can perform effectively against maneuvering threats.

4.2 Prediction Considerations

Fire control systems may be further classified by the sophistication of their prediction schemes, varying from manually introduced lead to non-linear prediction. For this discussion, prediction schemes will be restricted to first and second order processes. Apparent velocity multiplied by projectile time of flight defines first order prediction and second order prediction adds to this value the product of apparent acceleration and one half the projectile time of flight squared.

It has been seen in Fig. 2.7 that large variations of apparent velocity and acceleration occur for maneuvering vehicles moving on serpentine paths. These variations have adverse consequences on first order or linear predictors but may be exploited when second order predictors are used.

A detailed inspection of these variations in apparent velocity and acceleration are shown in Fig. 2.9. The resulting apparent motion is attenuated compared to a simple harmonic motion model. The apparent velocity is continuous, but has abrupt changes, the apparent acceleration has discontinuities and the cyclic frequency of the apparent motion is a multiple of the fundamental simple harmonic motion frequency which is dependent upon the ratio of the angular magnitude of the circular arc of the vehicle's path relative to a complete revolution of simple harmonic motion.

4.3 Example

To demonstrate the differences in performance of first and second order predictors for threat maneuvers on sinusoidal and serpentine paths the following example is given. A vehicle is assumed to be moving at a speed of 10 meters/sec on a 28-meter radius as shown on Fig. 2.8. The resulting radial acceleration is 3.5 meters/sec² and the cyclic period is 17.4 sec. This level of acceleration and velocity are representative of the performance characteristics of existing military vehicles. Assuming a projectile time of flight of $\frac{17.4}{8}$ sec or $\frac{1}{8}$ of the simple harmonic cyclic period and then calculating the lead error of different fire control systems engaging this evasive head-on vehicle, an envelope of performance can be determined for the first and second order predictors. Two specific firing times are assumed; when the apparent velocity is at a maximum value and when it is zero. Fig. 4.3 describes the lead errors resulting from this analysis. Another motion profile is considered where the evasive vehicle moves on a $+45^\circ$ arc of the original circle and then transfers to another similar circle having a reversed curvature similar to the illustration shown in Fig. 2.7. The lead errors for this serpentine maneuver are also listed in Fig. 4.3. Comparison of the two sets of data indicates a wide difference in performance of the first order predictor system for projectiles fired at the time of apparent maximum velocity (i.e., 7.40 meters and 1.74 meters for the serpentine and sinusoidal motions respectively). For the first order predictor a projectile fired at minimum velocity will have the same lead error of 8 meters for both serpentine and sinusoidal movement. Closer inspection of this difference in performance of first order lead systems engaging mobile/agile vehicles moving from point A to B will reveal that the serpentine model is more realistic than the simple harmonic motion model which does not permit the maneuvering vehicle to progress from point A to point B. When the serpentine evasive model argument is analyzed it is observed that the apparent velocity is constantly changing and never approaches a steady value as in the simple harmonic motion case, therefore, the apparent acceleration is never in the vicinity of zero for extended periods of time. This set of conditions persists throughout the apparent cycle of evasive vehicle motion. If the apparent acceleration is seldom in the vicinity of zero as it is for periods of the simple harmonic motion model, then this fact can be used to improve the prediction process. Incorporation of the apparent acceleration in the prediction process provides significant improvement, relative to the first order predictor, for the example being discussed. The lead error in the situation where the projectile

is fired at apparent zero velocity is the same in both the sine wave and serpentine maneuvers, .27 meters, which is much lower than the 8 meter error obtained for the first order predictor. For a projectile fired at maximum apparent velocity on the sinusoidal and serpentine paths the lead error is 1.75 meters and 1.54 meters respectively. A tabular presentation of these findings is also shown in Fig. 4.3 which shows the dramatic improvement that can be realized by developing a fire control system that has the ability to estimate apparent acceleration and combine it with flight time to develop a component of lead that is not attainable for a first order or linear predictor process.

5. SOURCES AND MAGNITUDES OF GUN POINTING ERRORS

5.1 Error Sources

If ballistic effects are not considered, miss distance is primarily a function of the tracking, estimation and prediction processes. Some understanding as to the relative distribution of the errors for each of these processes will be of interest. Fig. 5.1 shows an envelope of miss distances caused by tracking process errors for both manual and automatic tracking systems. Threats exhibiting maneuver levels up to 0.5g were considered for the generation of these data. Improvement in miss distance for auto track operation is dramatic as evidenced by the narrow band compared to the wide band for manual tracking. Other data, obtained from similar fire control systems having a manual track system, which included the error contribution of not just the tracking process but also the estimation and prediction processes indicates that the miss distance is much larger as shown in Fig. 5.1. For an engagement range of 1500 meters and assuming that the total miss distance is the root sum square of the individual miss distances, the error contribution of the estimation and prediction processes can be estimated from these data. If the total miss distance is bounded between 2.8 meters and 1.8 meters as shown on Fig. 5.1 and the tracking error for manual track is bounded between 1.3 meters and 0.8 meters, an estimator plus predictor error boundary between 1.15 meters and 2.7 meters exists. The ratio of the estimation and prediction errors to the tracking errors is between 70% and 350%. These calculations indicate that the estimation and prediction errors are at least as large as the tracking error and can be significantly larger. For the estimation and prediction errors just determined the total miss distance for an auto tracking system is calculated, assuming the miss distance contribution due to tracking error is 0.6 meters. The total miss distance for the auto track system is not significantly reduced from the manual

tracking system. This discussion implies that improvements in fire control system estimators and predictors are as acutely needed as the development of improved tracking systems. Not until the introduction of maneuvering threats are the basic limitations of the disturbed reticle fire control system fully appreciated in terms of the distribution of errors between the three processes. Improvement in the estimation process is needed to provide improved second order prediction capability. The essential requirement for improved estimation capability is small tracking errors. The inherent capability of a stabilized sight is generally accepted. The fact that the tracking loop in a stabilized sight tracker is a man-instrument servo system and not a man-turret servo system permits the fire control system designer to maximize tracking performance capability.

5.2 Estimator and Predictor Errors

In Reference [1] a series of maneuvering paths were used to determine the performance of different types of fire control system predictors. A declassified presentation of the results is shown in Figs. 5.2 and 5.3. Descriptions of the maneuvering vehicle paths which were used in the study are shown. A min-max envelope, similar to the tracking error envelope shown in Fig. 5.1 is used to describe the magnitude of the errors. Different ranges, reflecting different times of flight are included. The presentation is intended to illustrate the relative contribution of tracking and prediction errors and further demonstrate the significant improvements that are realized when the order of the predictor is changed from first to second. The prediction errors are subdivided into four experiment groups to demonstrate the influence of maneuvering level on the magnitude of the errors. Fig. 5.2 shows the small degradation caused by maneuvers not developing apparent accelerations in excess of $0.05g$. Also seen in Fig. 5.2 is the extreme prediction error degradation caused by a vehicle developing peak apparent accelerations of 6 meters/sec^2 while executing cyclic oscillations of $1/8 \text{ Hz}$. The improvement obtained by using a second order prediction capability is significant. Fig. 5.3 shows errors caused by both the estimation and prediction processes for a conventional fixed gain estimator. Compared to the error caused by a first order predictor only it is seen that the estimator contributes an additional source of error. The design methodology alternatives used to estimate apparent motion are discussed in a later section of this paper. Plots of the performance index of the different predictor levels and fixed gain estimator designs is shown in Fig. 5.4 for both long and medium range engagements. For the medium range engagement the second order predictor is not

significantly degraded. The fixed gain first order estimator is adversely influenced by increased accelerations. For the long range engagements, the predictors degrade rapidly until the acceleration approaches 3.5 meters/sec^2 . Thereafter, a relatively fixed degraded level of performance is realized. For the longer range engagements the knee of the performance index for the first and second order predictor processes is realized within the acceleration envelopes of tests that have been conducted with existing tactical vehicles. High mobility and agility capability as being anticipated in forthcoming test programs is not required to demonstrate the inadequacy of existing fire control systems. This seems to augment the case for obtaining improved estimator/predictor processes for fire control system performance improvement if these levels of mobility and agility are to be considered as threats.

6. FIRING DOCTRINE

The prediction example discussed in Section 4.3 illustrates the potential advantages of firing the projectile at a maneuvering threat when the apparent velocity is at a minimum value. It may be possible to exploit this situation by designing a fire control system that mechanized this concept. Precedence for this approach currently exists within the training programs for manual systems. The so called "ambush" tactic calls for generation of a fixed lead coupled with firing of the round as the minimum presented area of the tank is seen. This "ambush" procedure can be interpreted as being a crude version of a second order prediction, where the lead angle is one half time of flight squared multiplied by the apparent acceleration. As seen in Fig. 2.9 the apparent acceleration is nearly constant within a relatively large part of the arc of the serpentine maneuver. The lead angle or offset is not too sensitive to the apparent velocity, which is near zero during this time period. The maneuvering vehicle time constants that control this acceleration, which is aligned normal to the tracks or wheels of the vehicle, insure that the direction of vehicle movement is not altered once curvilinear motion has commenced. Knowledge of the magnitude and direction of the vehicle's velocity and acceleration with respect to the line of sight can be combined to provide an automated second order predictor that automatically combines maneuver characteristics with optimum firing doctrine. Threshold levels on acceleration can be used to select an option to use either first or second order prediction. This adaptive feature would ensure that performance against essentially non-maneuvering threats would not be adversely influenced by higher order prediction systems.

7. MECHANIZATION OF FIRE CONTROL SYSTEMS

7.1 Classical Fixed Gain Lag-Lead Filter

The classical method of filtering a noisy signal such as the handle bar movement in a gun fire control system designs a single input - single output lag-lead filter which both attenuates and phase shifts the tracking input signal. In applications where the noise or handle bar jitter is sufficiently removed in frequency from the tracking data related to the maneuvering threat, this method attenuates the handle bar noise and only slightly attenuates and phase shifts the desired line of sight rate signal. The output of the filter or estimator is the smoothed line of sight rate that is combined with the time of flight in the first order predictor. The gain of this filter or estimator is fixed, which means that the band width of the fire control system is independent of the maneuvering level of the threat. Fig. 7.1 shows the signal flow input-output for this type of filter and Fig. 7.2 shows the band pass characteristics of such a filter.

7.2 Sub-Optimal Adaptive Estimator

The possibility of obtaining improved fire control systems by using multi-variable, sub-optimal estimators in place of fixed gain filters suggests that the design of such a system should be initiated [2]. The following discussion will focus on the problem of threat modeling to illustrate the geometric and computational considerations involved in the sub-optimal methodology. The utilization of all available information describing the threat is the most important distinguishing feature that exists between classical fixed gain filters and multi-variable-sub-optimal estimators. Fig. 7.1 shows the input-output signal flow for for such an estimator. Emphasis is placed on the development of a mathematical model of events that occur in the real world. For a maneuvering threat-engaging fire control system the model may be separated in two parts, with each part having a deterministic and random description. The formulation of the sub-optimal estimator requires that an analytical description of both threat movement and tracking sensor operation is contained in the structure of the data processing algorithm. Requirements for the statistical characteristics of both the threat movement and tracking sensor can be intelligently derived from engineering data. The incorporation of these descriptors into the overall structure of the sub-optimal design filter methodology is summarized in Fig. 7.1. Fig. 7.3 is a block diagram of the real world-mathematical model interface used in this approach.

The orientation of the maneuvering threat with respect to the line-of-sight is constantly changing for a maneuvering vehicle. These geometric effects can be exploited when sub-optimal filtering methodology is used. Vehicle movements are characterized by mobility parameters defined in the body axes of the maneuvering vehicle. In particular, accelerations produced can be described in terms of vehicle mobility and agility. These vehicle oriented accelerations can be rotated into the tracking sensor line-of-sight coordinate frame thereby providing an adaptive feature to the filtering process as shown in Fig. 7.1. The band width of the filter is adjusted according to the maneuver level of the threat vehicle. The ability of the sub-optimal estimator to perform effectively is related to the validity of the structure of the mathematical model. In reality there cannot be a perfect match between real world events and mathematical representations in the estimator and this leads to the concept of a sub-optimal design. It is this time varying performance that is correlated to the target maneuvering which permits the sub-optimal estimator to provide better estimates of the state of the apparent threat movement than can be obtained from a fixed gain filter design.

8. SUMMARY

An analytical description of a maneuvering threat vehicle has been developed to study the degradation of fire control systems performance. The fundamental processes of tracking, estimation and prediction have been identified and related to the three generic types of fire control systems currently in existence; manual, disturbed reticle and stabilized sight-director systems. The relative contribution of each of the three processes to miss distance has been discussed and the results of a simulation describing the sensitivity of the order of the prediction processes to the miss distance is presented. Classical design methodology is compared to sub-optimal design methodology and the performance improvements obtained from second order or acceleration predictors is discussed.

9. REFERENCES

- [1] H. H. Burke, "An Investigation of the Performance of Gun Fire Control Systems Engaging Maneuvering Threats", AMSAA TR 234, August 1978.
- [2] H. H. Burke, T. R. Perkins, J. F. Leathrum, "State Estimation of Maneuvering Vehicles Via Kalman Filtering", AMSAA TR 186, October 1976.

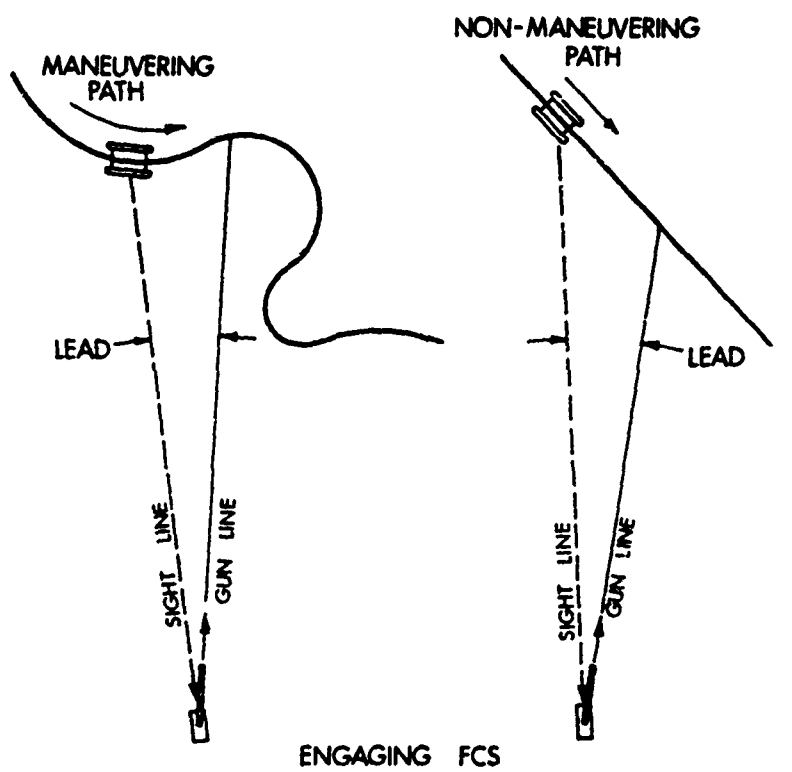


Figure 21 Fire Control System and Threat Movement.

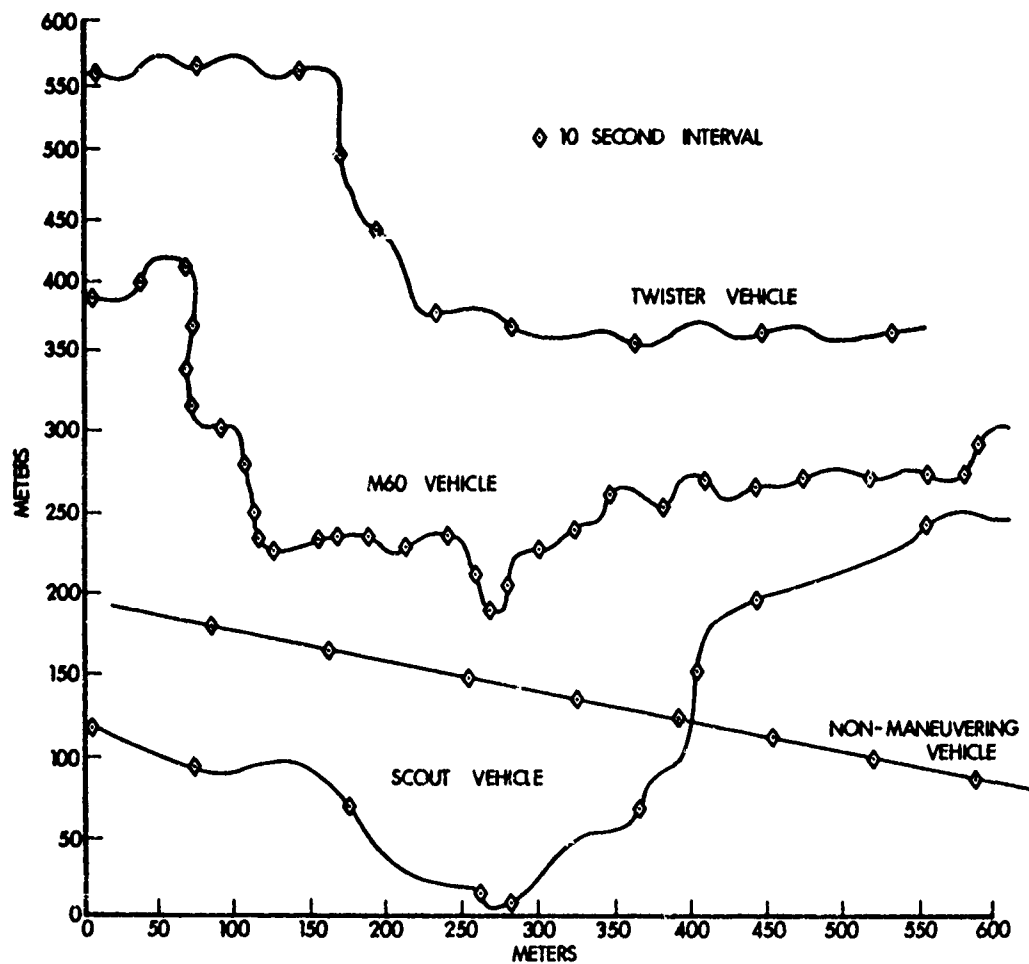


Figure 2.2 Empirical Maneuvering Vehicle Paths.

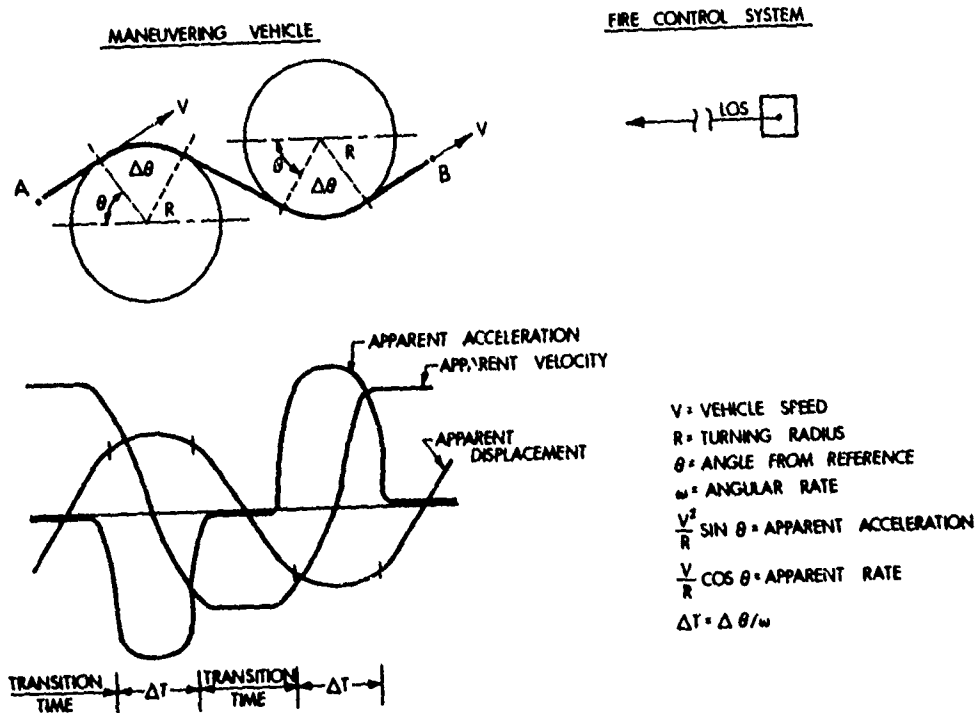


Figure 2.7 Intentional Maneuvers and Apparent Vehicle Movement.

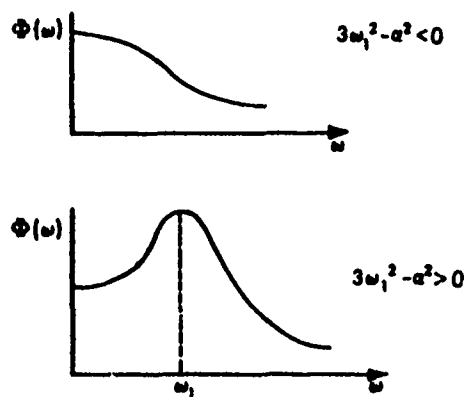


Figure 2.6 Curves of Power Spectral Density.

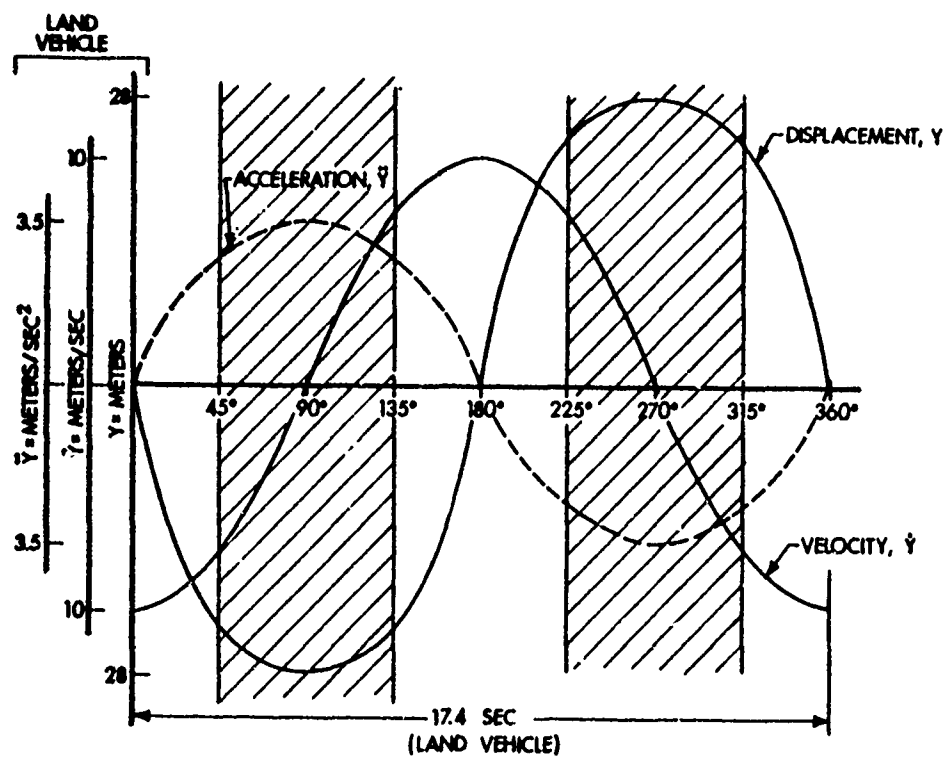


Figure 2.8 Sinusoidal Motion — Along Y Axis Showing 90° Arcs of Vehicle Movement.

2.9 LAND VEHICLE SERPENTINE MOTION

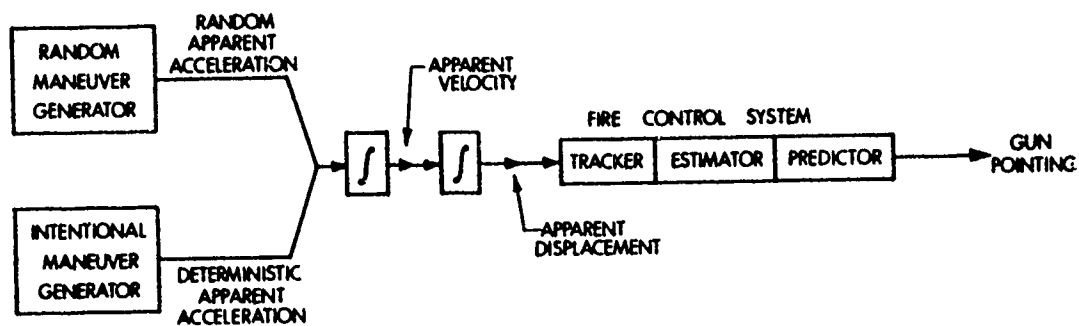
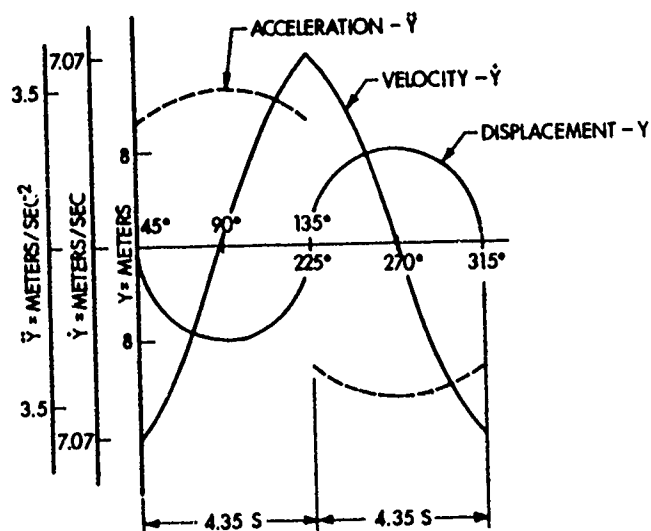


Figure 2.10 Total Analytic Maneuver.

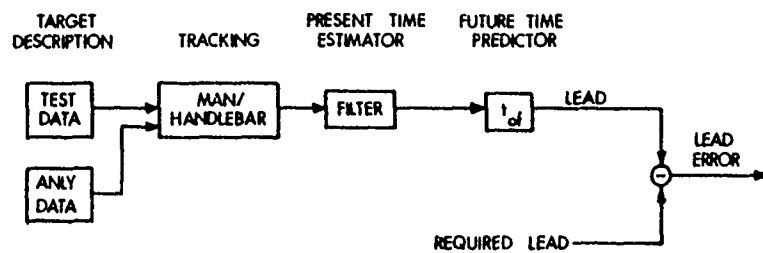


Figure 3.1 Block Diagram of Predictive Fire Control System.

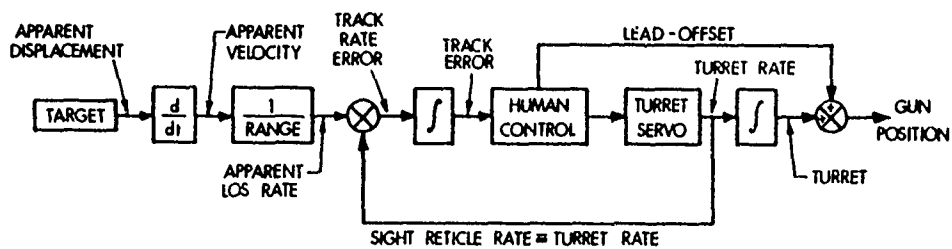


Figure 4.1a Manual Track and Lead System.

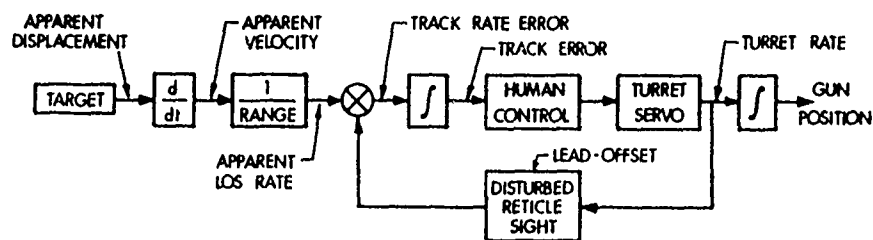


Figure 4.1b Manual Track and Disturbed Reticle Sight Lead System.

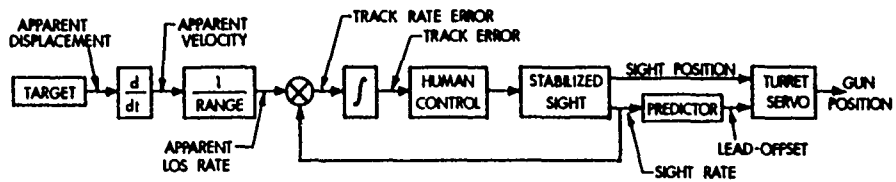


Figure 4.1c Manual Track and Stabilized Sight-Director System.

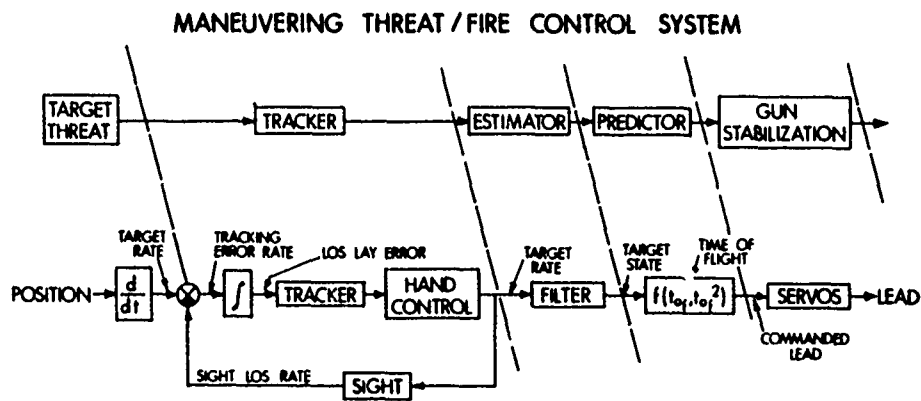


Figure 4.2 Stabilized Sight-Director FCS.

MOVEMENT MODEL	FIRING	LEAD ERROR (M)	
		1st ORDER	2nd ORDER
SINUSOID	V _{MAX}	1.74	1.75
	V _{MIN}	8.00	.27
SERPENTINE	V _{MAX}	7.40	1.54
	V _{MIN}	8.00	.27

$$t_{of} = \frac{17.4}{8} = 2.175 \text{ SEC}$$

Figure 4.3 Lead Errors Resulting from Prediction.

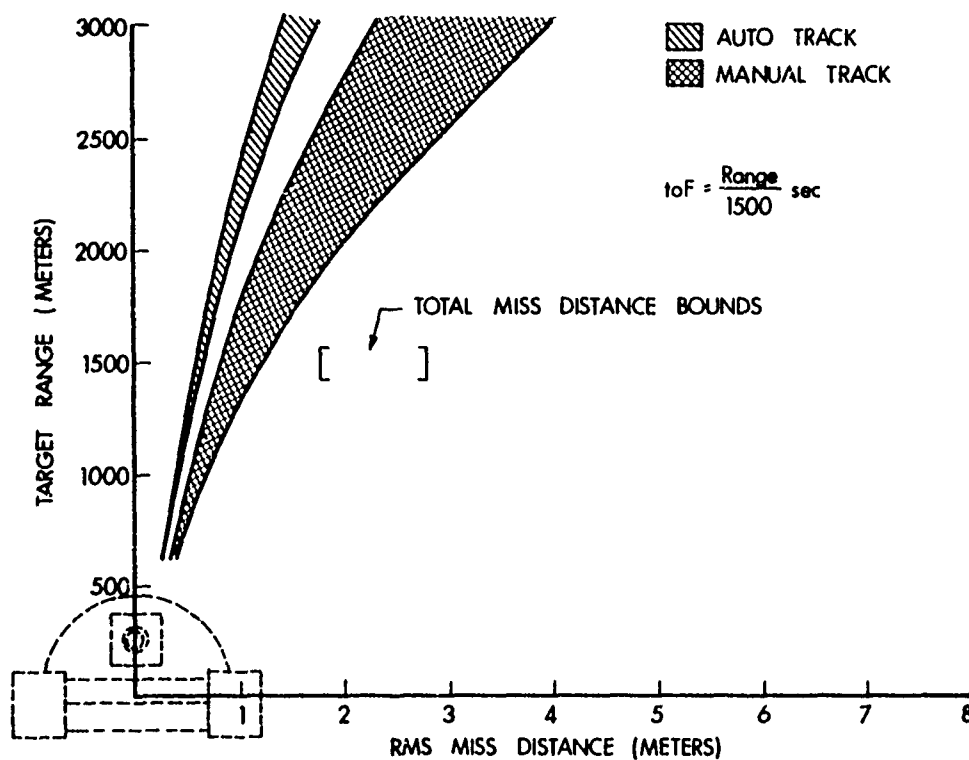


Figure 5.1 Miss Distance Due to Tracking Errors.

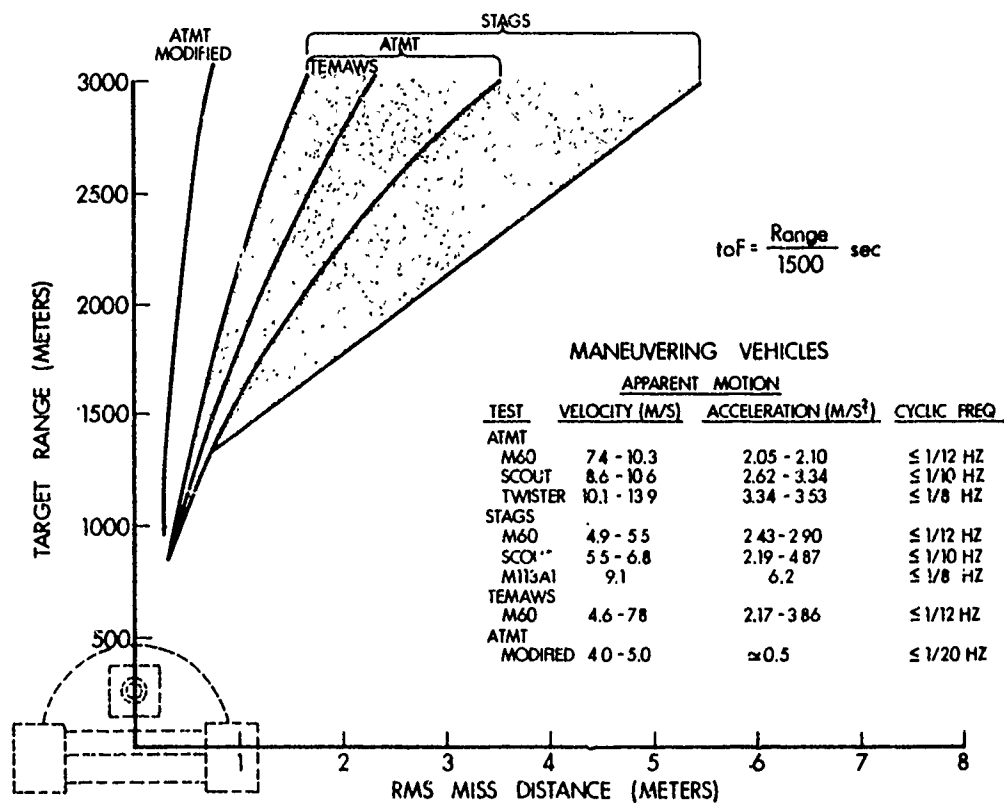


Figure 5.2a Miss Distance Due to 1st Order Prediction Errors.

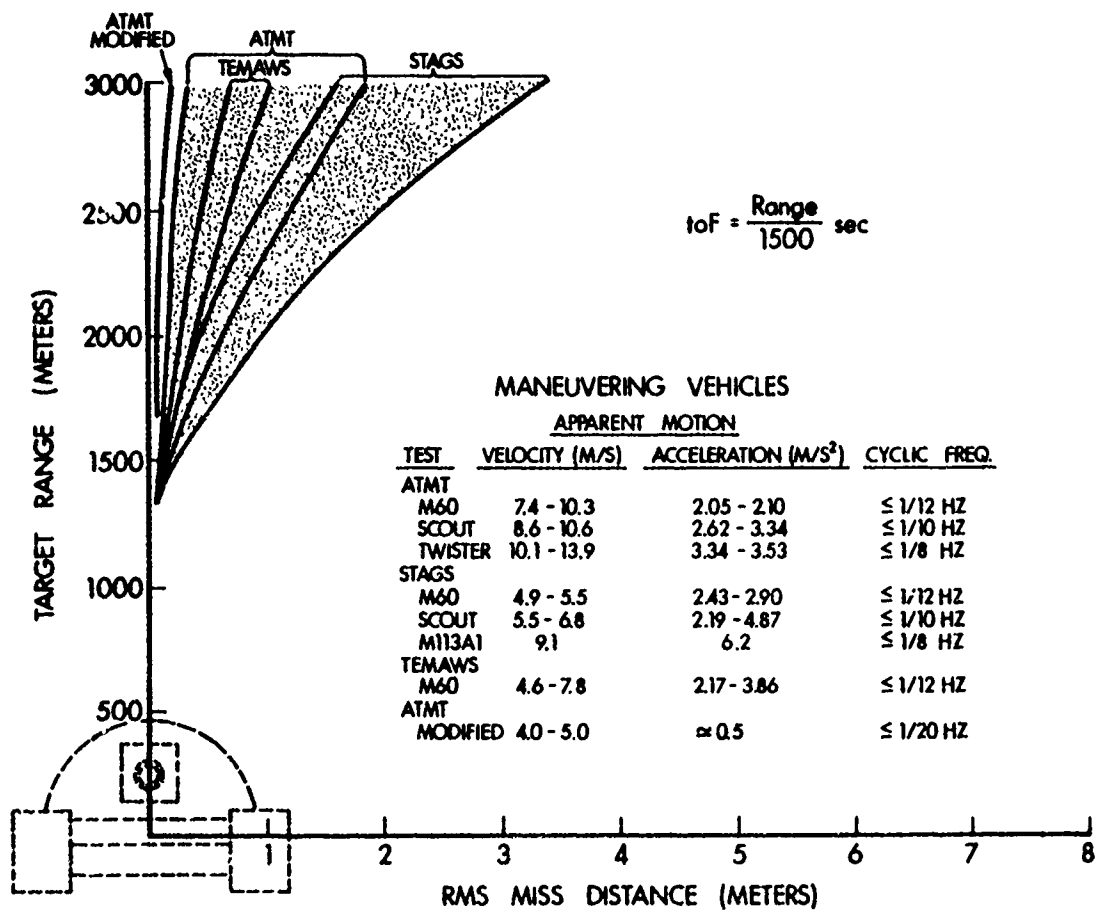


Figure 5.2b Miss Distance Due to 2nd Order Prediction Errors.

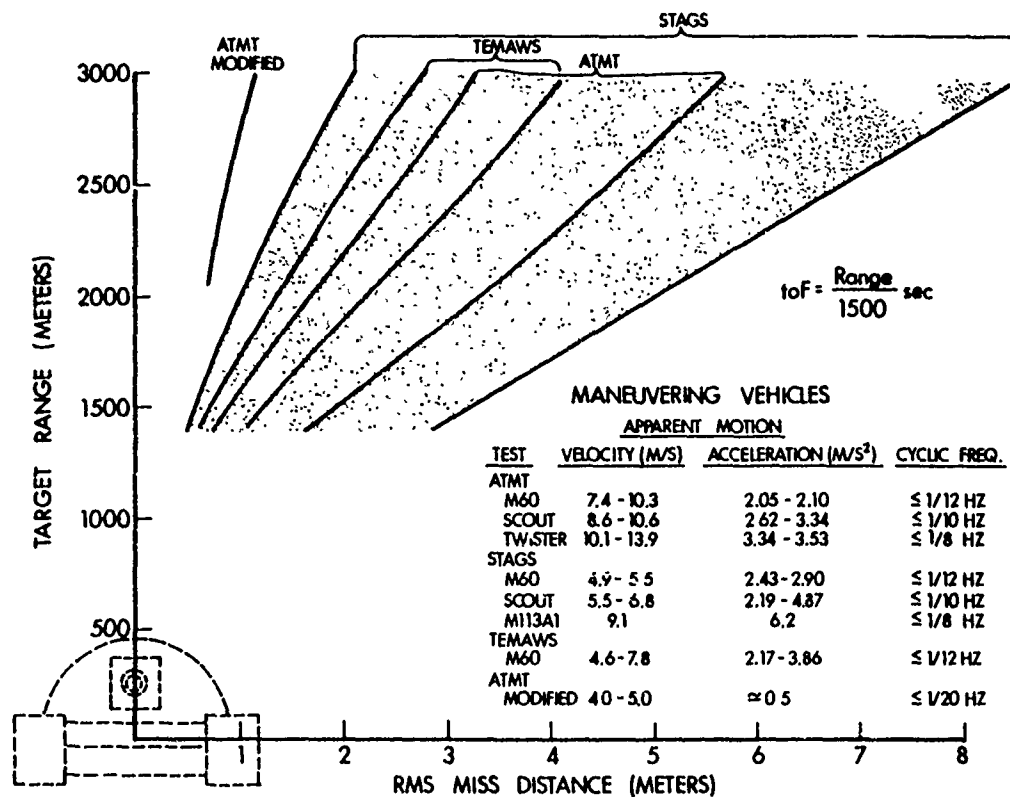


Figure 5.3 Miss Distance Due to Fixed Gain Estimator Plus 1st Order Prediction Errors.

M60, SCOUT, TWISTER VEHICLES
MEDIUM ENGAGEMENT RANGE

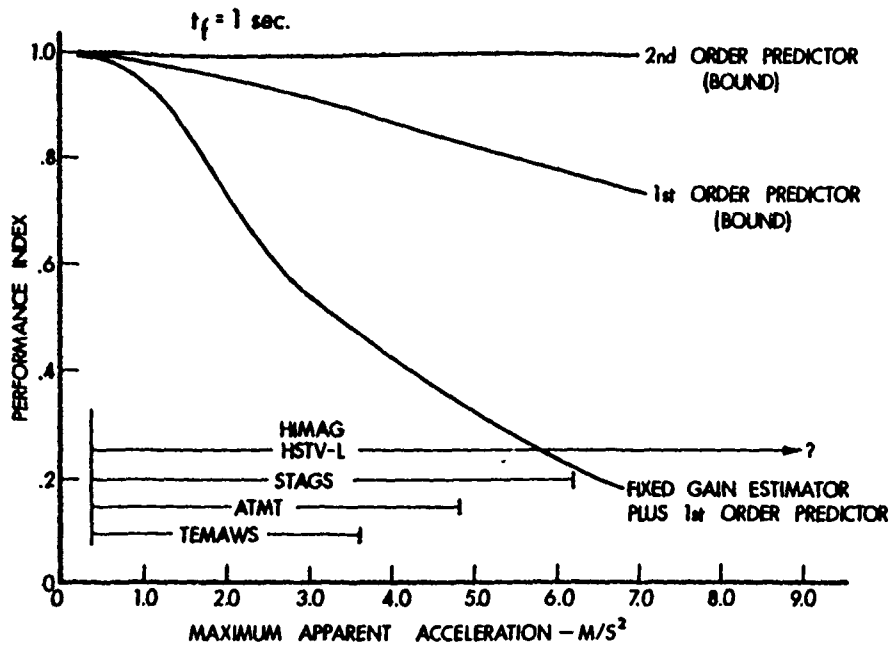


Figure 5.4a Summary of Fire Control Systems Relative Performance.

M60, SCOUT, TWISTER VEHICLES

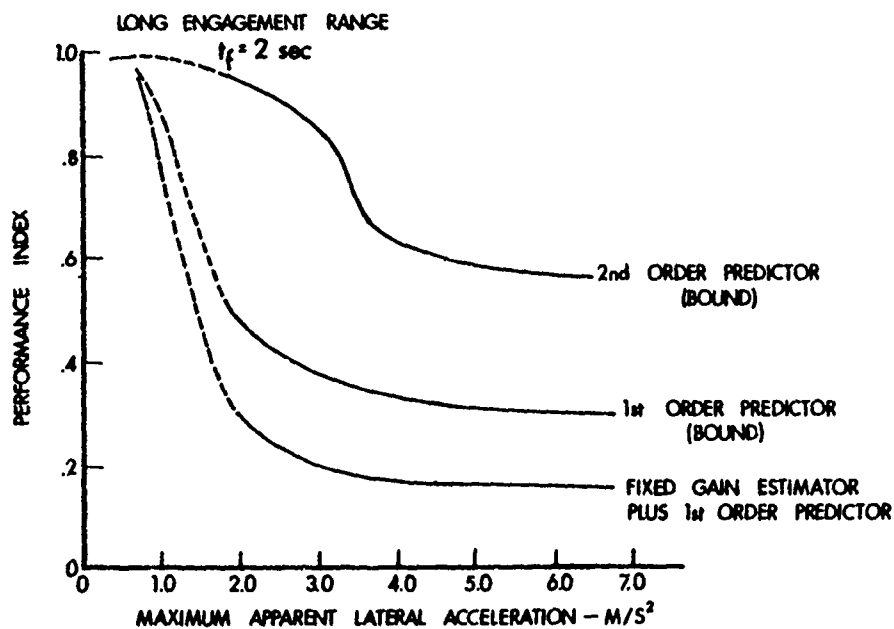


Figure 5.4b Relative Performance of Fire Control Systems.

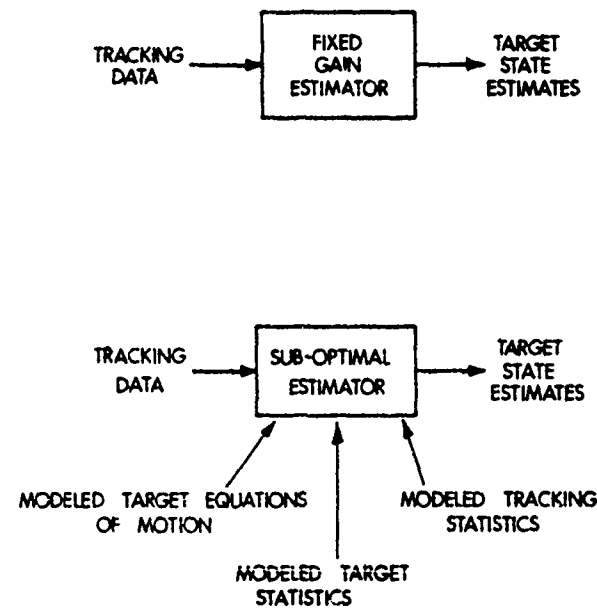


Figure 7.1 Comparison of Fixed Gain and Sub-Optimal Estimators

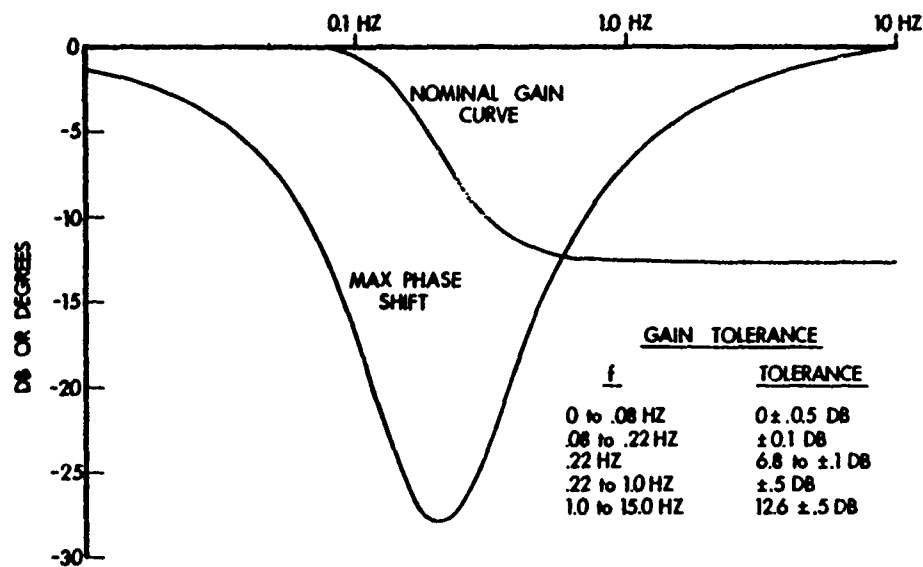


Figure 7.2 Gain/Phase Characteristics of Fixed Gain Filter.

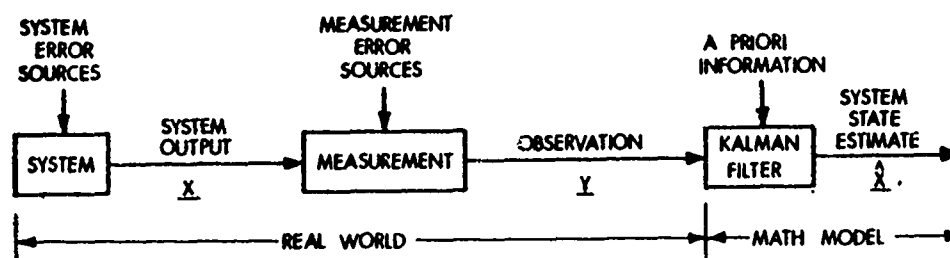


Figure 7.3 Block Diagram Showing System, Measurement and Kalman Filter.

NAVAL FORCE STRUCTURE PLANNING

Leonard P. Gollobin, President

Presearch Incorporated
Arlington, VA 22202 USA

ABSTRACT. A practical approach is presented for naval force structure analysis which recognizes the important policy/analyst interfaces at the beginning and end of the planning process. Specific computation techniques are not given, but concepts and methods of analysis are shown, together with examples of intermediate and final results. What is stressed is the logical breakdown of the force structure planning problem from broad policy and scenario considerations, down into sub problems which are analyzable by conventional military operations techniques. Then results are consolidated into force structure alternatives useful to and understood by policy level decisionmakers.

This approach was used for the quantitative analysis in the recent U.S. Navy SEAPLAN 2000 Force Structure Study for which the author was technical director. Some force option summaries from this study are shown to illustrate the output.

1. INTRODUCTION

1.1 Naval force structure planning is a logical process that links national, defense and navy policymakers, and naval analysts. This paper presents some views about how such planning can be accomplished in a practical way, and draws upon the experience of the author, especially the recent SEAPLAN 2000 Study performed in the United States for the Secretary of the Navy.

1.2 First the hierarchy of the force structure analysis is presented. Then a means is given for breaking down broad requirements for naval forces into analyzable parts, and then recombining them into force structure alternatives. Each stage of the analysis is then discussed in practical terms.

2. FORCE PLANNING ANALYSIS

2.1 Planning Hierarchy

2.1.1 Naval force planning begins with identification of the missions and roles of the navy in the accomplishment of national objectives. The loop is closed when the decision-maker is presented with the predicted results for alternative force structures, for accomplishing the same national objectives. Some of the steps illustrating the ensuing logic are shown in Figure 1. In the figure, the downward flow signifies segregation of the variables, while the upward flow represents the aggregating, or consolidation phase of the analysis. In the latter, discrete one-on-one engagement results are pulled together to ultimately reflect the predicted outcome of the conflict.

2.1.2 Viewed from the top down, the force structuring process breaks down high level national scenarios and naval missions into analytically manageable campaigns, operations, group and finally unit engagements. Along the way, each step is characterized by more detailed models, inputs, assumptions and results. From the bottom up, results from models at each level have to be consolidated into tactical, strategic, scenario and fiscal options understood, and able to be acted upon, by the highest policy levels. Here the mass of detailed results and trends have to be joined together into meaningful (usually more simplified) form to be useful.

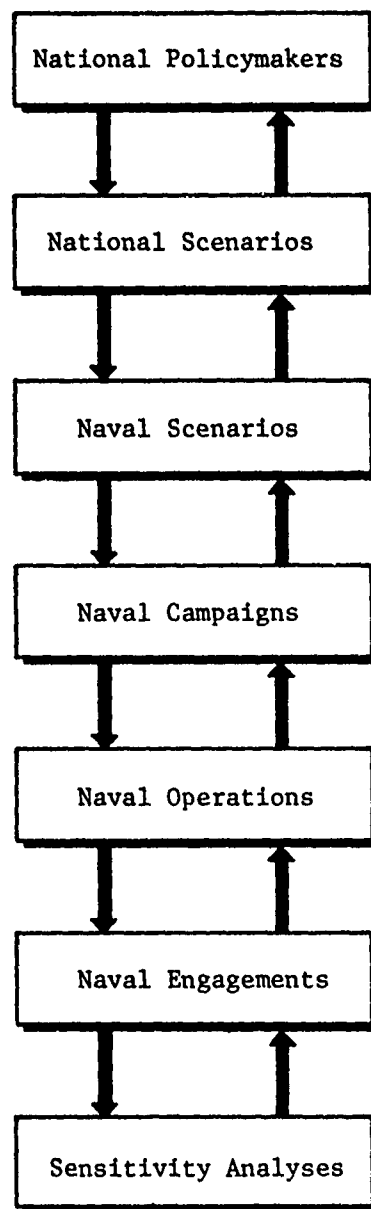


Figure 1
Naval Force Structure Analysis Hierarchy

2.2 Broad Approach

2.2.1 The basic principles used in the quantitative analysis recognize that there are force structures appropriate to different national policies and naval strategies, and the total force in each case is the sum of all the parts. These parts should account for: (1) all the roles and missions for which the navy is assumed responsible; (2) time phasing, which considers transit times and the sequencing developed in the scenarios that reflect strategies and tactics for both sides; (3) units out of action or unavailable as the result of operational attrition (such as mechanical failures), losses to enemy action, and scheduled maintenance, overhaul or modernization; and (4) assets which provide support to other navy forces (including fixed installations).

2.2.2 The force needed to accomplish a set of objectives, such as maintaining the sea lines of communication (SLOC), begins with forces in being at the start of crisis or conflict; takes credit for the generation rate of new (or activated reserve) forces over the period of the conflict; and the residual force required to be on hand at the cessation of hostilities in sufficient numbers to satisfy national criteria of "winning."

2.3 Scenarios

2.3.1 The critical politico-military interfaces, or linkages in force planning occur at the beginning and end of the analysis process. At the beginning, scenarios are created which describe the types of conflicts in which naval forces are likely to be engaged; the agreed-upon threat; and roles and missions for the navy in support of national policies. At the end, the quantitative analysis will relate the ability of alternative forces to accomplish these objectives, and at what cost, so that the policy maker has a basis for allocating his resources.

2.3.2 The results of analysis are frequently driven by the assumptions made; scenarios are no different. Hence the scenarios have to be drawn even more carefully than some of the detailed inputs for the later quantitative studies, since scenario assumptions will spread through the entire analysis. An overwhelming level of detail should be avoided at the scenario level, since such detailing tends to minimize flexibility and creativity in force options and usage. However, sufficient detail is needed that adequately describes: (1) a realistic and believed chain of politico-military events; (2) own and enemy order-of-battle;

(3) deployments of own and enemy forces; (4) the time frame and time schedule of the events, and (5) significant details such as warning and tactical surprise in a buildup spanning peacetime, contingency and wartime conditions. Backing up the scenarios is a detailed threat assessment that identifies specific platforms, combat suites, weapons and supporting technical characteristics that are inputs to effectiveness models.

2.4 Depicting the Conflict

2.4.1 The next step is to develop naval campaigns and operations within the context of the scenario, much along the lines of preparing war plans, or laying out war games. It is important that both sides utilize their assets in the best way to accomplish their own objectives, and so it is frequently useful to designate own and enemy tacticians and strategists as members of the analysis team.

2.4.2 The identification of campaigns involves breaking down the scenario along geographic or time phased lines, such as Western Region Defense, SLOC Protection, Home Defense, Retaliatory Phase, which are readily identified by the non-military policy maker. Operations are linked to roles for naval forces (mine countermeasures, convoy escort, shore bombardment against bases) as the determinant of the types of engagements expected during the conflict. An illustrative breakdown is shown in Table 1.

2.4.3 The outcome of this part of the planning process is a family of detailed force movements and warfighting sequences; timetables for the action, and identification of encounters between opposing forces at the task group or one-on-one level. A tabulation of these encounters provides a checklist of engagement models (detection, closure, attack, reattack, etc.) needed for each area of naval warfare [antisubmarine (ASW), antiair (AAW), antisurface (ASUW), mine, amphibious, air strikes, harbor defense--]. The complexity of the conflict and forces involved also indicate the method (manual or machine) and form of the accounting model used to aggregate, time phase and generally keep track of the results of lower level engagements.

2.5 Engagement Analysis

2.5.1 The tie-in between engagement analysis and the overall force planning hierarchy is illustrated in Figure 2. Some examples of combat systems and measures of effectiveness

TABLE 1
ILLUSTRATIVE NAVAL PORTION
OF NATIONAL CONFLICT

<u>Campaigns</u>	<u>Operations</u>
SLOC East	<ul style="list-style-type: none"> ● ASW patrol ● ASW escort ● Surface patrol ● Contact prosecution ● Mine countermeasures ● Support
SLOC West	<ul style="list-style-type: none"> ● ASW escort ● Mine countermeasures ● Protect fishing fleet ● Support
Homeland Defense	<ul style="list-style-type: none"> ● Defensive mining ● Coastal patrols ● Support
Protective Strikes	<ul style="list-style-type: none"> ● Protective reaction strikes against naval bases ● Naval gunfire support ● Amphibious operations ● Support

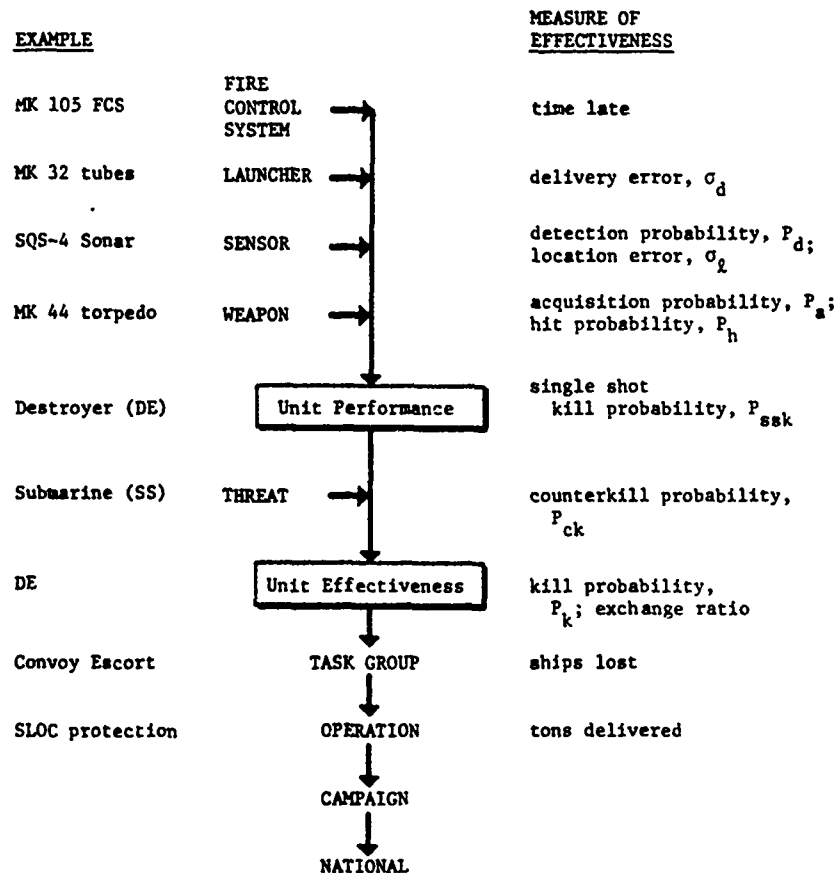


FIGURE 2
ENGAGEMENT ANALYSIS HIERARCHY
(ASW Example)

are shown for an ASW-oriented example. Each element of the combat suite supplies its unique entry to the performance models, which when combined with tactical inputs and the behavior and characteristics of the target, yield estimates of unit effectiveness. These can be combined to predict task group effectiveness.

2.5.2 Using the illustration of Figure 2 which describes a convoy escort, it is the destroyer combat suite that provides the building blocks for computing overall SLOC protection effectiveness. Starting with the basic threat inputs on submarine population and distribution from the scenario, the frequency or likelihood of occurrence of submarine encounters is established, for example from barrier models. Conversely, the probability that a submarine is detected by convoy escort screens is given by acoustic search models, assuming here for simplicity that sonar is the only available sensor. (In a real case, all sensors on all platforms that could provide alerting and detecting events, would be considered).

2.5.3 The events from detection through classification, localization, attack, reattack, target hit and kill are depicted as conditional probability terms in a weapon system effectiveness (WSE) model. The inputs to such models may conveniently be obtained from generalized relationships of the type and form shown in Figure 3, which in this case is from a statistical, "cookie-cutter" model applied to a large kill radius weapon dropped against a submarine. The advantage of generating a data base of such parametric relationships, is that as weapons, threats, etc. change, the impact of such changes can be read immediately.

2.5.4 Counterkill by the target is considered in a similar WSE model, and final results can be expressed in the form of exchange ratios where the combatants can both be attrited in an engagement. The logic flow for such a model is illustrated for a submarine/destroyer engagement in Figure 4. The models can range from simple, expected value types, with aggregated, statistical inputs derived theoretically or from at-sea tests, extending all the way to complex, very detailed simulations that address all of the operational, technical and environmental factors simultaneously.

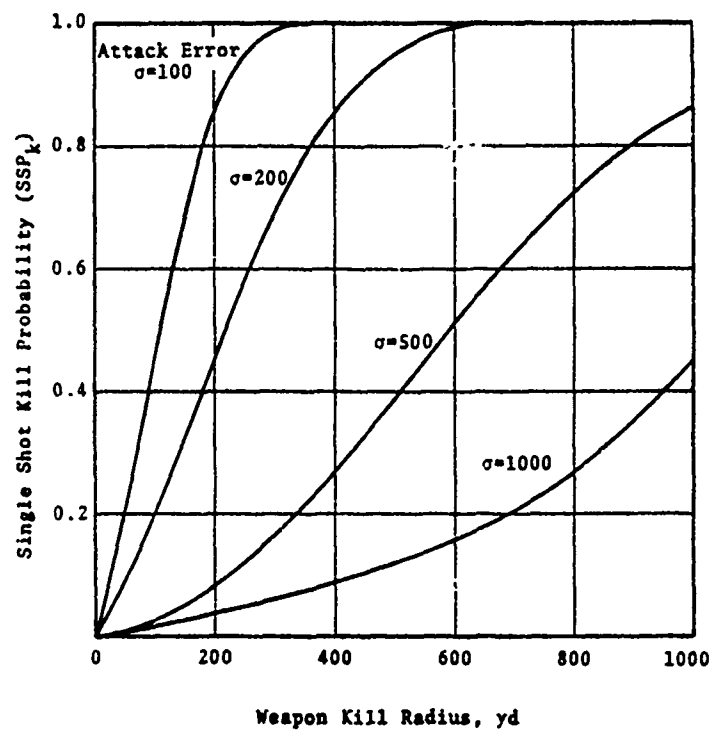


FIGURE 3
SINGLE SHOT KILL PROBABILITY
VERSUS KILL RADIUS AND
ATTACK ERROR

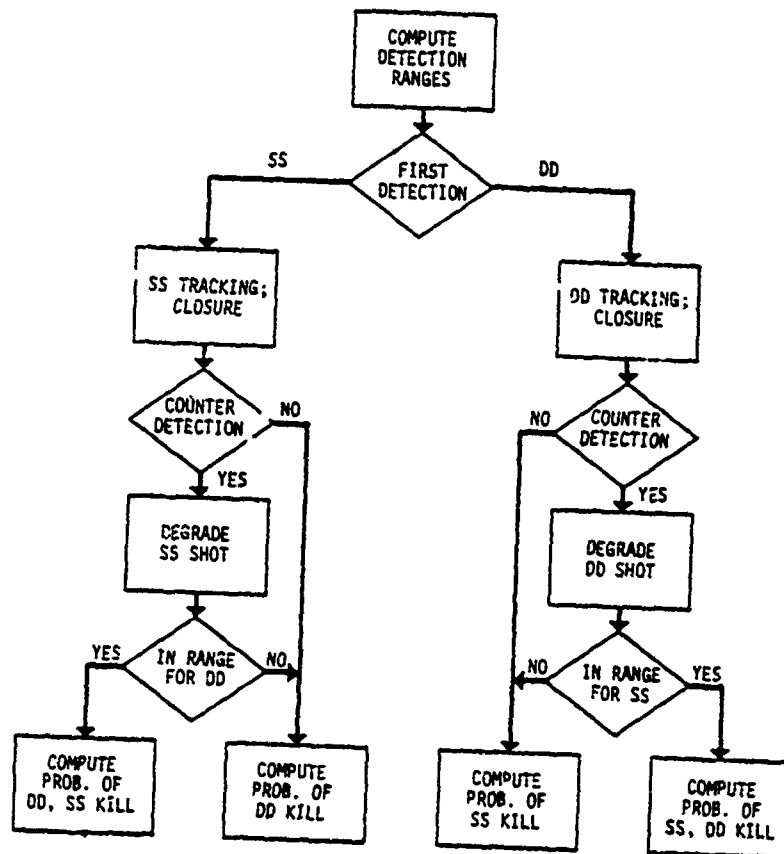


FIGURE 4
SIMPLIFIED FLOW DIAGRAM FOR DD/SS ENGAGEMENT

2.5.5 It is worthwhile to comment that the credibility and accuracy of the results of the analysis are not necessarily related to the complexity and detailing of the models, but are more likely to be governed by the quality and realism of the assumptions and numerical inputs. The writer has also observed that frequently too much of the initial analysis effort goes into model development, or preparation of inputs and utilization of very complex models, whose level of detail may exceed the quality of rough technical or cost input data. It is believed valuable to start with a more balanced, first-order approach which quickly establishes principal trends, key variables, critical assumptions or gaps in the data, the magnitude and form of the results, and identifies areas needing and justifying further, more detailed examination.

2.5.6 Along similar lines, it is usually more efficient and meaningful to select a "baseline" case and run it through the entire analytical process, before attempting to formulate and execute a large number of parallel cases. By selecting mid range, "normal" characteristics for the variables in the baseline, it is possible to rapidly develop a good feel for the number and extent of excursions that need to be tested to satisfy a meaningful matrix of results. These can often take the form of sensitivity checks, with relatively few variables perturbed at one time, and which provide very good insight into the trends of the results. These sensitivity studies are appropriate for examining the effects of such changes as tactics employed by both sides; alternative combat suites; threat characteristics and the impact of new technology on sensors, weapons and platforms. The analysis of these perturbations, in turn, is the basis for the alternatives and cause and effect relationships that are presented to the policy-level decision-maker.

2.6 Costing

2.6.1 Many of the decisions relating to naval force size and mix involve consideration of the cost to buy and operate naval assets. This writer has observed that the quality and refinement of effectiveness analysis often far exceeds that of cost analyses, despite the fact that a decision may be made on cost-benefit grounds, or sometimes with offsetting technical factors, on the basis of cost alone. It is a disservice therefore to combine eloquent analytical results with very rough cost estimates,

claiming an eloquent resultant. Here again, the baseline case concept applied to costing, provides a good feel for where to emphasize detailed costing.

2.6.2 Since costing is usually regarded as a "softer" area of analysis, typically with large uncertainties, it follows that independent checks are helpful for increasing confidence in the results. For example, generating independent "top down" and "bottoms up" costs provide management- and technically-oriented views, as well as numerical comparisons of the same item. (Top down costing usually refers to estimates developed from broad gauge cost factors, indices, prior production or extrapolation from related programs. Bottoms up costing usually refers to item-by-item, piece/parts costing, in which all of the cost components are identified and summed).

2.6.3 Because the manner in which budgets may be drawn affects the decision process, several kinds of cost estimates are frequently needed. The initial buy, or acquisition cost, is of obvious interest since it represents a near-term event, and is a first order guide to the relative expense of alternatives. Life cycle costs (LCC) take into account operating and support costs and the useful life of an asset, and may include the cost of modernizing or updating to improve effectiveness and extend the useful life. Unless there are significant differences between alternatives in useful life or the other long-term cost items, LCC tends to smear differences between alternatives. This is especially true when the real issue may be more sensitive to the near-term availability of funding.

2.6.4 While the main thrust of this paper is not costing, a checklist of life cycle cost items is included for convenience:

- purchase of hardware, including spares, test and peculiar support equipment
- management of the procurement process, including preparation of drawings, specifications, and purchasing and standardization efforts
- research, development and test of new systems, components or interfacing equipment

- operating costs, including manpower, fuel, expendable items, repair and maintenance
- support costs, including training, base operations and logistic support
- modernization and upgrading of existing systems.

3. PRESENTING THE RESULTS

3.1 Approach

3.1.1 Going back to the aggregating process in the first figure, the task is to bring together unrelated encounter results in minesweeping, ASW, amphibious warfare, etc., into naval mission/national objectives terms. A pattern for consolidating these results is suggested below, whose end result is to display what different force levels, mixes or capabilities can accomplish, and at what cost. Some intermediate results are illustrated for the hypothetical conflict shown in Table 1, and some final presentations are illustrated by borrowing from SEAPLAN 2000 examples.

3.1.2 The steps involved in consolidating results of the analysis are given below and discussed in the following paragraphs.

- Unit and task group engagement results are correlated against their MOEs as force size or capability vs. results achieved and losses suffered
- Each event that employs naval forces is identified, and the force required to accomplish a defined objective (e.g., 80% SLOC throughput) is computed for each event
- Time phase the events, and compute forces needed for all events comprising a naval operation
- Similarly, time phase naval operations and compute forces needed for all operations comprising a campaign
- Account for losses and unavailability and provide for support forces

- Combine campaign results for each scenario
- Generate force alternatives and compare them in terms of effectiveness in meeting missions and objectives, and at what cost.

3.2 Examples and Discussion

3.2.1 One form in which one-on-one and group engagement results can be consolidated, and which provides a parametric data base for excursions, is shown in Figure 5. The upper ASW curves relate convoy escort force size to protected ship losses under different threat conditions, and the lower curves relate mine countermeasure force size to ship losses to mines. Results can be scaled for different combat suites or mine types, on the same axes. These curves can be used to determine what force capabilities are needed to satisfy objectives such as achieving an 80% throughput in the SLOC (which may be based on policy and military grounds).

3.2.2 If naval forces were not multimission, and time phasing or location prevented units from handling more than one task over the period of conflict, then the total force needed would be the sum of units required to perform all of the tasks independently. But time phasing frequently allows the same forces to be used for different tasks separated in time, allowing of course for transit times to shift locations, if needed. An example at the naval operation level is the repeated use of the same destroyer for convoy escort, except as limited by simultaneous convoys. A case at the campaign level might be shifting minesweepers from SLOC defense after harbors are cleared, to homeland defense coastal patrol in later stages of the conflict.

3.2.3 An example of force sizing for a campaign broken down by separate naval operations, is given in Table 2. Numbers of ships, boats and aircraft required to accomplish patrol, escort, etc., objectives are shown, first computed as independent events, and then after taking credit for time phasing. By way of illustration, suppose that a total of six submarine contacts needed to be prosecuted by aircraft over the period of conflict, but that they were so spaced in time that they could be handled by two aircraft. Then

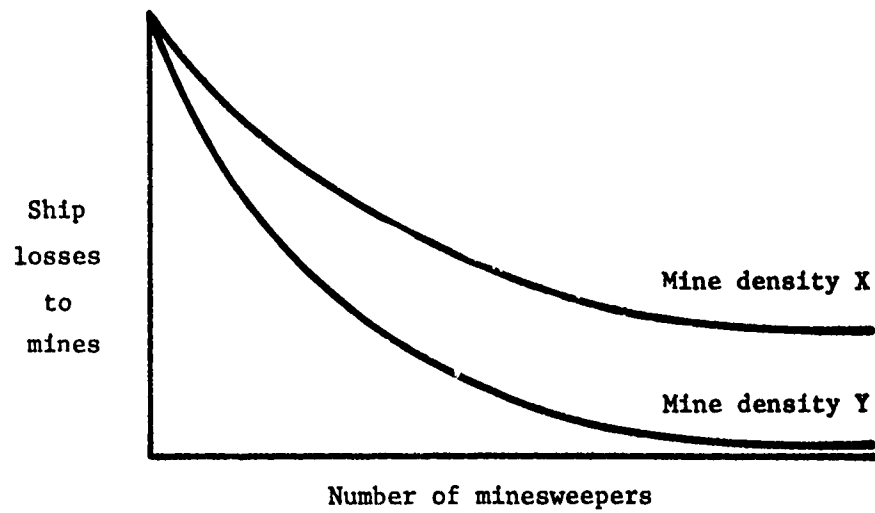
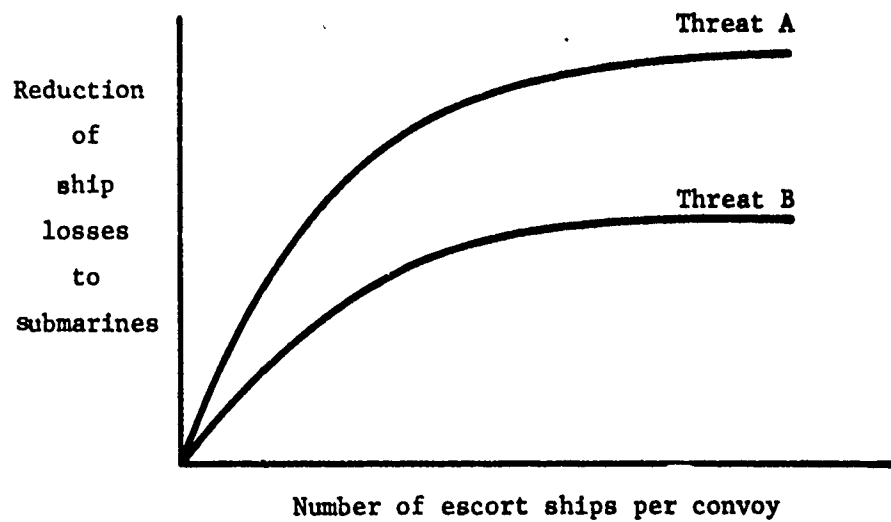


FIGURE 5
TYPICAL AGGREGATED ENGAGEMENT RESULTS

TABLE 2
ILLUSTRATIVE CONSOLIDATION OF ENGAGEMENT
RESULTS, SLOC EAST CAMPAIGN

Operation	Numbers of Platforms Required					
	Independent Events			With Time Phasing		
	Ships	Boats	A/C	Ships	Boats	A/C
ASW Patrol	3	-	10	2	-	6
ASW Escort	4	-	-	3	-	-
Surface Patrol	4	20	4	3	12	2
Contact Prosecution	4	-	6	2	-	2
Mine Countermeasures	6	-	-	4	-	-
Support	6	10	-	4	6	-
Sub-Totals	27	30	20	18	18	10

the indicated force requirement would be two aircraft, as shown in the table. A simple accounting model is used to keep track of all forces. (Even for the complex SEAPLAN 2000 analysis, a manual force accounting model was found adequate).

3.2.4 Table 3 illustrates the campaign level of aggregation, introducing the concept of scenario variations, and providing for forces that are unavailable. (The latter applies to ships in overhaul or operationally unavailable. Losses to enemy action are included as part of the basic force estimates derived from the engagement analyses). This example has been extended to show what increments are required over an assumed available baseline force in being, together with the estimated cost for procuring the force increments.

3.2.5 Some interesting examples of an approach to succinctly summarizing the results of a major, detailed force analysis

are given in the SEAPLAN 2000 Summary, ^{1/} and are repeated here. Table 4 shows alternative force structures developed from worldwide force allocations, building up from combat suite and unit performance estimates and one-on-one engagement analyses just as described above. The ships shown are "notional," that is, characteristics for actual classes have been normalized to some baseline performance. Ship-building costs to meet the force levels shown for each option in the year 2000 are given in Table 5 for the Five Year Defense Plan (FYDP) period, which is the usual planning cycle for the U.S. Department of Defense.

3.2.6 The linkage between the force structures in Table 4 and the accomplishment of national objectives is illustrated in Table 6. In this case, the matrix entries reflect analyses of force requirements under a wide range of scenarios generated using Defense and Navy guidelines. The value of this presentation has been stated to be that it bounds the problem from a peacetime through full scale war climate, and provides a very concise, bottom line assessment of the alternatives.

^{1/} U.S. Department of the Navy, SEAPLAN 2000, Executive Summary, UNCLASSIFIED (March 28, 1978).

TABLE 3
ILLUSTRATIVE CONSOLIDATION OF CAMPAIGN
FORCE NEEDS AND COSTS
(results time phased)

Campaign	Scenario # 1			Scenario # 2		
	Ships	Boats	A/C	Ships	Boats	A/C
SLOC East	18	18	10	6	8	6
SLOC West	10	12	6	10	12	6
Homeland Defense	2	24	2	4	20	4
Protective Strikes	8	30	*	12	30	*
Sub-Total	38	84	18	32	70	16
Unavailable at 15%	6	13	3	5	11	2
Totals	44	97	21	37	81	18
Δ Force Added to Baseline	9	27	3	2	11	0
Δ Procurement Cost	450	54	9	100	22	0
\$79M	513			122		

* Would utilize non-navy assets

TABLE 4

SEAPLAN 2000: ILLUSTRATIVE OPTIONS FOR YEAR 2000 NAVY

SHIPS	OPTION 1	OPTION 2	OPTION 3
SSBN	25	25	25
CV	10	12	14
SSN	80	94	98
AEGIS SHIPS	10	24	28
SURFACE COMBATANTS	210	252	272
AMPHIBIOUS SHIPS	52	66	78
OTHER	52	64	70
TOTAL	439	535	585

TABLE 5

SEAPLAN 2000: ILLUSTRATIVE SHIPBUILDING COSTS
(FY 79 \$B)

	FY 79	FY 80	FY 81	FY 82	FY 83	79-83 Average
Option 1	4.7	5.9	6.5	6.3	5.5	5.78
Option 2	4.8	7.6	8.2	7.8	8.9	7.46
Option 3	4.8	7.6	8.6	8.6	9.2	7.75

TABLE 6
COMPARISON OF SEAPLAN 2000 FORCE OPTIONS

MEASURE	OPTION 1	OPTION 2	OPTION 3
MAINTAIN STABILITY	<ul style="list-style-type: none"> • RELAX CURRENT FORWARD DEPLOYMENT • ADVERSE PERCEPTIONS OF POWER 	<ul style="list-style-type: none"> • MAINTAIN CURRENT DEPLOYMENT • RESOLVE VERSUS SOVIET GROWTH 	<ul style="list-style-type: none"> • CURRENT DEPLOYMENT AT IMPROVED ROTATION RATE • ENHANCED PERCEPTION
CONTAIN CRISES	<ul style="list-style-type: none"> • CRISIS/DEPLOYMENT TRADEOFF • HIGH D-DAY SHOOTOUT LOSS 	<ul style="list-style-type: none"> • SUSTAIN FORWARD DEPLOYMENTS DURING A CRISIS • CREATE SAGs 	<ul style="list-style-type: none"> • SUSTAIN FORWARD DEPLOYMENTS DURING CRISES • SIGNIFICANT RESIDUALS
DETER GLOBAL WAR	<ul style="list-style-type: none"> • SOME SLOCs • NO FORWARD OPS AT BEST • DEFENSIVE 	<ul style="list-style-type: none"> • PROTECTS SLOCs • ENABLES 2-4 FORWARD OPS • SECOND FRONT OPTION 	<ul style="list-style-type: none"> • ALL-AROUND SUPERIORITY
RISK ASSESSMENT	HIGH RISK; MINIMAL CAPABILITY; NOT FLEXIBLE	MINIMUM ACCEPTABLE RISK; MAINTAINS SELECTIVE SUPERIORITY VS. SOVIETS	LOWER RISK; PROVIDES HEDGE AND OPTIONS

4. CONCLUSION

Naval force structure planning is a synthesis of qualitative and quantitative analysis. The policy linkage is that different scenarios and campaigns qualitatively define the job to be done by naval forces, while the quantitative analysis measures the ability and cost of alternative force structures to do these jobs. The success of this planning process depends on close coupling between policy makers and analysts first in setting objectives, and then in understanding and assessing the options.

A LOGISTICS REQUIREMENTS STUDY

RONALD C. RUSH
LIEUTENANT COLONEL, USAF

Logistics Staff Officer
Deputy Chief of Staff - Systems and Logistics
Pentagon, Washington, D.C.

ABSTRACT. This paper presents an analytical model of logistics requirements necessary to support a theater war scenario. These requirements include medical, materiel, personnel replacements and transportation and are generated from the employment of friendly military forces and equipment against a postulated enemy threat. This model analyzes the individual phases of the war scenario and generates logistics requirements by type of class/unit, location within the battle area, supply points for distribution of materiel, and transportation modes for movement of the materiel. Medical casualties by type and transportation capabilities by type of unit are also identified. All logistics requirements are capable of being identified by variable time phases of the war.

1. INTRODUCTION

1.1. Overview

The Simulation and Gaming Methods for Analysis of Logistics (SIGMALOG) System is a current logistics simulation model which identifies theater logistics requirements. SIGMALOG is a total logistics analysis system in that it is capable of handling supply and transportation requirements for all classes of material. At the same time, being a computerized set of models, individual segments of the system can be analyzed in depth.

1.2. SIGMALOG Structure

SIGMALOG has two basic parts, SIGMALOG I and SIGMALOG II. SIGMALOG I provides an analysis of the intra-theater requirements necessary to support an employment of forces. SIGMALOG II compares requirements generated by SIGMALOG I with the assets available and the capabilities of the intra-and inter-theater transportation systems. Our application is only addressing SIGMALOG I at this time.

1.3. SIGMALOG Phases

The employment of the SIGMALOG System progresses through three major phases - preparation (presimulations), simulation and post simulation. The presimulation phase includes defining the scenario; determining the method of operation and the conditions and assumptions which must be included; the collection and identification of doctrinal, conceptional, intelligence, and operational guidance; and the acquisition of required data. The simulation phase involves the preparation of necessary input data in prescribed formats, the running of the computer models and the analysis of model reports. Any simulation may require the recycling of the entire scenario or rerunning of selected models until acceptable results are attained i.e., the optimum practical employment of manpower, material and facilities. The post-simulation phase involves preparation and publication of the report.

2. SYSTEM PARAMETERS

2.1. Regions

Employment of the SIGMALOG System, whether in total or in

part, requires the definition of the basic parameters. The geographical area in which the military operation is postulated must be described in sufficient detail to permit identification of significant tactical and logistical operations at key points. This is facilitated by dividing it into subareas, called "regions", in which these key points are located. The forward regions on the forward edge of the battle area (FEBA) are occupied by tactical units such as combat units and armored brigades. To the rear of the forward area, regions are established in the combat zone to represent higher echelons in the command and support structure. In the communications zone (COMMZ), regions are normally established for each major logistical complex. Airfields and seaports, wherever located, may be included in the above regions or designated separate regions.

2.2. Time Periods

The time span of the military operation being simulated must be divided into segments, called "time periods", which will permit the post-simulation of tactical operations and the examination of logistical operations at significant periods of time during the operation. An initial time period - time period 0 - is used to represent a moment in time immediately preceding the initiation of combat operations, which are assumed to begin on D-day.

2.3. Roles

The Forces included in the scenario have been identified as a function or role, in order to associate types of units by role to varying levels of activity resulting in more detailed logistics support requirements. Typical roles for Army units are combat, air defense, combat support and combat service support. Role codes for Prisoners of War (POW) and Refugees can also be analyzed. Therefore, in a forward region, as an example, during a single five day time period, combat units could be in intense combat, air defense units could be operating at a reduced level of activity, combat support units could be operating at a normal level of activity and combat service support units could be operating at a reserve level of activity. The levels of activity, i.e., intense, normal, reduced and reserve are input against each of the units in the applicable regions so that their variations in activity can be measured in the amount of material requirements generated.

3. SYSTEM DESCRIPTION

The components of SIGMLOG I are a set of data base programs and logistics-function simulating computer-assisted models as follows: (See Fig. 1)

3.1. Force Employment Data Automation (FEDA) Programs

These are a series of ADP routines developed to provide automatically certain portions of the large volumes of input data required by the Data Base Programs (DBP) and the Force Employment Model (FEM). The FEDA Programs reduce significantly the manual coding and verification processes. In addition, they provide the data more efficiently and timely.

3.2. Data Base Programs (DBP) I - IV

These routines accumulate information from three sources, reformat these data elements to provide capability with the processing procedures of the entire system, and copy these data on magnetic tapes. The restructured data serve as input to certain other models in the system. The FEDA--provided troop list is an initial source to the DBP. Among other Army automated files which serve as a second input source are the Computerized Movement Planning and Status System (COMPASS) File, the TOE Master File, and the DCSLOG Data Processing Center (DDPC) Unit Weight File. The third input source is manually input card data.

3.3 Force Employment Model (FEM)

The FEM simulates the time phased deployment and employment of forces and the variable postures of combat and combat support forces. Produced from this simulation are certain related workloads, troop lists, and strength aggregations. These provide input to the subsequently-processed functional models. The FEM programs require data from FEDA and DBP magnetic tape files and manually-prepared cards. The tape files contain information defining characteristics of the troop units and their equipments. Policies must be provided for the deployment and employment of troop units, among which are specific time on position and location of employment. Routing rules must define theater arrival time and intra-theater routing. Further, the model simulates activities of additional groups of personnel. Refugee and Prisoner of War (POW) factors are

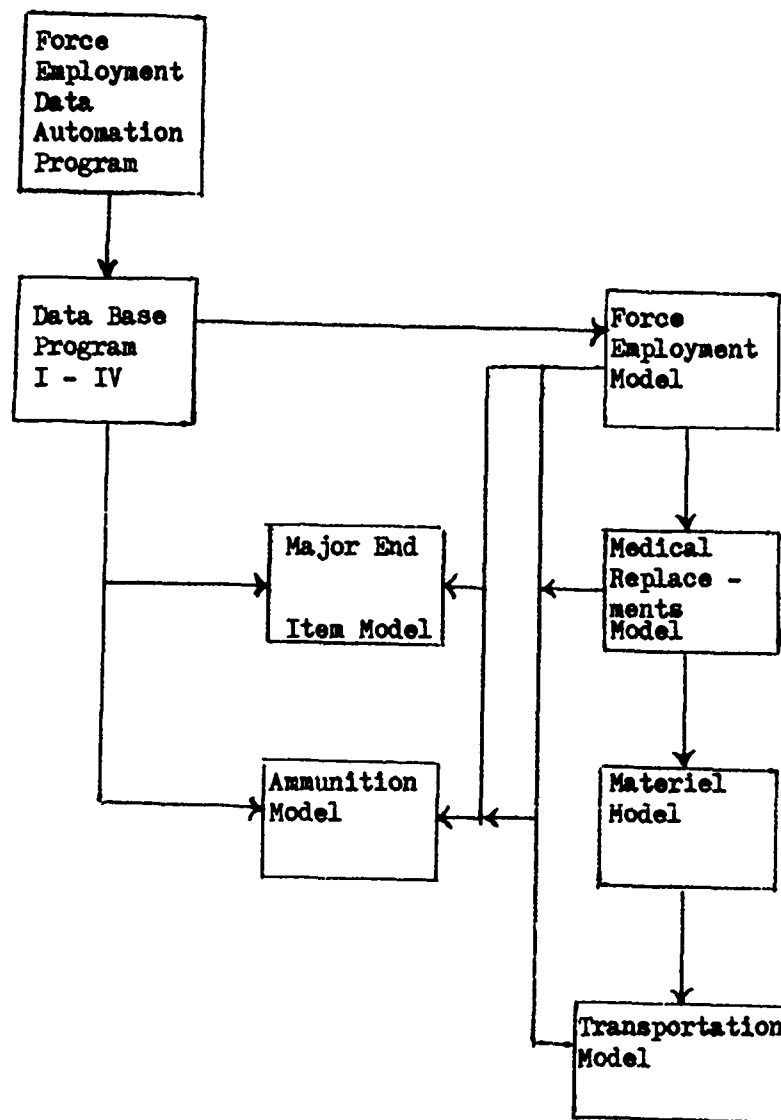


FIGURE 1 SIGNALLOG FLOW CHART

included particularly because of the significant effect these activities have on material requirements. Combat unit peaks can be identified in the form of total manpower available due to mobilization of reserve units and movement of active units. Forward regions would normally include artillery, tank and motor type units during early time periods. Air Defense units would be actively engaged during the initial periods of a war when opposing forces would be striving for air superiority. The analyzation of the resultant data from this model will provide an overall assessment of the wartime status of all units at any time period.

3.4. Medical/Replacements Model

The Medical Model computes workload imposed on the medical system while supporting theater military operations. Echelon workloads are presented in terms of hospital beds required and distribution of patients through evacuation, returned to duty, and death. Time-phased personnel replacements for all non-returned to duty, killed in action, and missing in action are required. Such quantities determine the impact of receipt, processing, and intra-theater movement on the Material and Transportation Models. Troop-strength information is received from a FEM tape file. Medical policies pertaining to hospitalization and evacuation and casualty factors are card input. Output reports provide medical requirements in terms of fixed and non-fixed hospital beds. The Material Model calculates material support for patients occupying beds. Regional time-phased replacements are generated by the Replacements routines. Gross values, sums of the wounded in action, disease and non-battle injured, killed in action, and missing in action are reduced by returnees to duty.

3.4.1. Wounded In Action (WIA)

Depending on the scenario being used, the majority of the WIA usually occur in the front most regions along the FEBA during the initial periods of a war.

3.4.2. Killed In Action (KIA)

Similar to the WIA statistics, the overwhelming majority of the KIA usually occur in the front most regions during the early time periods.

3.4.3. Disease and Non-Battle Injury (D&NBI)

The D&NBI casualties in the forward regions are generally not significantly greater than those in the rear regions. In fact some of the regions other than the front-most regions to the FEBA may show increases in D&NBI casualties since non-battle injuries could increase as units move to the rear most region.

3.4.4. Missing In Action (MIA)

The majority of the MIA'S occur in the front-most regions in the initial periods of a war.

In analyzing the overall casualties the one interesting statistic that stands out is that the total D&NBI casualties almost always equal or exceed the total casualties generated by the other three categories (KIA, WIA and MIA). Hospital bed requirements would be generated by time period as the casualties occur. The personnel replacements required are generated by the model to meet the casualties that were killed, wounded or missing as well as those D&NBI casualties that were not returned to duty.

3.5. Material Model

3.5.1. Inplace Stocks

This model quantifies selected operational aspects of a theater supply system. Requirements are receipt, storage, and shipment of material. Upon these material tonnages are based a significant portion of the intra-theater transportation requirements and the facilities needed to store the supplies. Tape file input from the Force Employment Model, modified by Medical/Replacements values, contains troop strengths to be supported in theater regions for appropriate time period intervals. Troop strength is classified by "user" a term denoting an aggregation of personnel consuming material at a common rate. Input card data define user consumption rates in pounds per man per day, according to resupply and stock-level policies. The supply system utilizes a chain of supply nodes, signifying supply-handling activities or depots, at which specified levels of material are received, stored and shipped. These sites are connected by transportation links over which supplies in specific quantities are moved. Inplace stocks

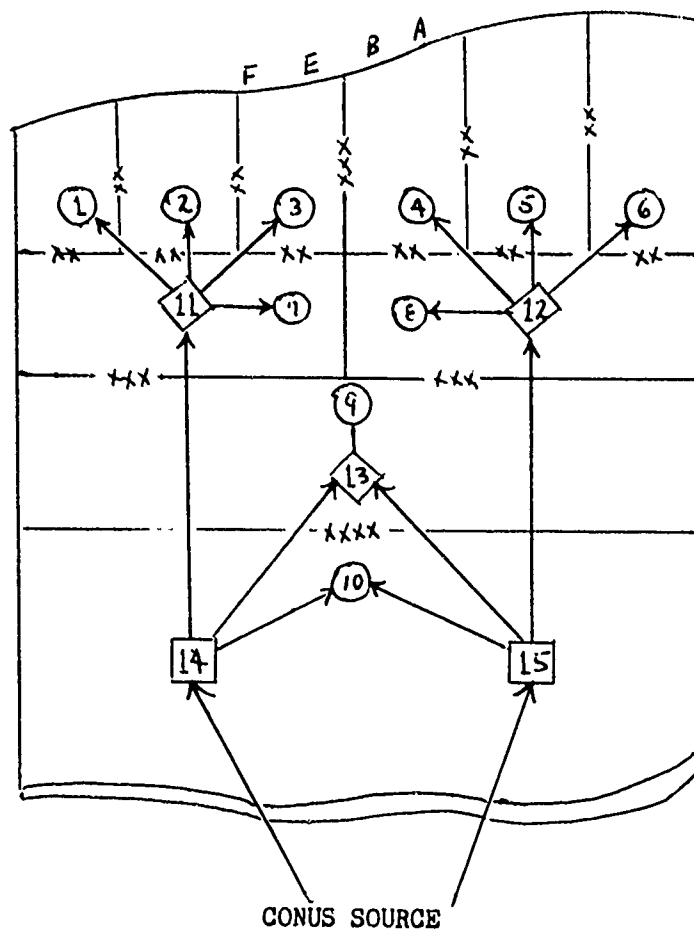
are established at these supply nodes at the beginning of time period 0.

3.5.2. Resupply Levels

Resupply begins when the material requirements exceed the available stock level. Initially customer demands are met from prescribed and basic loads of supply available at the unit level. Exhaustion of these quantities creates a demand for replenishment, which is met by pulling the proper amounts from the direct-support node. This replenishment action causes the node to examine its stock status, i.e., stock level, in relation to its computed stockage objective. The stockage objective is the computed stock level the node must attain by the next time period increment. If the stock level has fallen below the stockage objective, a replenishment action is generated. It is transmitted to the next supporting node in the chain which satisfies the demand, examines its own stock level in terms of its stockage objective, and so on through the defined system. The last node (s) in the theater chain pulls from the CONUS source to attain its required stock level. See Fig. 2. Material is considered in the model by supply class or subclass category. All 10 DOD Defined classes of supply can be used or these supply classes can be sub divided into as many as 20 different categories. Resupply activities calculated at each node include the number of short tons by material received, stored, and shipped, by category and time period interval. Shipments of material over the resupply links account for major transportation loads.

3.5.3. Operation

Forward supply nodes are usually class oriented such as Ammunition Supply Points (ASP's), POL Supply Points (PSP's) or Direct Support Units (DSU's). Supporting these supply nodes are General Support Units (GSU's) and finally in the rear areas the Depots support the GSU's and DSU's. To illustrate the use of this model, we will address the mechanics of how one class of material is portrayed, class V - ground ammunition, for example. The model output generates short tons of ammunition consumed at each supply node within each corps area. Additional ammunition consumption would be generated at rear areas by security forces. The largest portion of the ammunition would be consumed in the front regions at those forward supply



- Direct Support Unit
- ◇ General Support Unit
- Depot

FIGURE 2 SUPPLY NODE DISTRIBUTION

nodes (in this model they represent Ammunition Supply Points). The shipments to replenish issued material are generally issued from the closest intermediate depot or General Support Unit which stocks required ammunition. Each of the supply nodes are identified in a resupply chain for each class of material. Through the analysis of this data it can be determined in which time period and at which supply nodes the initial stocks will be depleted to the point of requiring resupply. By knowing this requirement in advance we can better prepare the material and transportation systems to respond to anticipated requirements. Also by continually comparing the stockage objective for each class of material with its stock level, we can program for a resupply of material from the CONUS supply channels.

3.6 Major Item Resupply Model

This model determines the time-phased major item losses and resupply of selected line item numbers (LIN) of equipment for the deployed force, and the percentages of air and surface movement of these required assets to the theater. Inplace stocks are input at the beginning of time period 0. Input data include parameters defining loss rates stock levels, percent of air shipment rates, etc. Selected authorized LIN densities are calculated from data contained on DBP and FEM files. Applying loss rates to these data produce replacement requirements. Another routine computes prestock levels and time-phased stockage objectives. Resupply calculations relative to each inter-nodal movement link and each nodal storage point are performed. End results are time-phased movements of LIN'S lost by the user, quantities to be moved between nodes and to be stored at the nodes, and total quantities to be transported inter-theater by air and sea in order to maintain stock levels. For the active weapons; i.e., howitzers, long range guns, etc., the replacement factor is relatively high because of the weapon systems proximity to the FEBA. Since the losses are directly proportional to the units activity, once full mobilization has occurred the loss rate remains fairly constant. By analyzation of the loss rate of equipment, operational and logistics planners can determine when the equipment loss will affect a combat units capability to fight. Based on equipment losses, resupply would be initiated upon receipt of CONUS shipment by sea except for critical Class VII items which would be airlifted into the theater during the early stages of a war.

3.7. Ammunition Resupply Model

A basis for determination of time-phased resupply requirements for selected items of ammunition, identified by DOD Ammunition Code (DODAC), is available through this model. Calculations are dependent on weapon density, tactical activity, expenditure and loss rate data, and stockage levels. Ammunition policies and factor modification are input by weapon system subset and by user code. These factors are translated by the program to the DODAC associated with each subset and user code. Tape data from FEM files provide troop unit information and an equipment file from one of the DBPs provides LIN authorization per troop unit. Ammunition is considered in rounds by type, fuses, charges, propellants, primers, grenades, etc. Options also exist to apply variable expenditure rates as a result of differing intensities of combat and loss rates for ammunition intransit and in storage.

The Ammunition Model in most applications examines the critical weapon systems. For Army units, for example, this might include weapons such as: M-16, 105 and 155 howitzer, 175 gun, machine guns, helicopter weapon sub-systems, etc. The model input for each of the weapon systems is in rounds per weapon per day. This gives better accuracy to the output since it is weapon systems oriented. The weapons density provided from the Data Base Program is given a rounds per weapon per day factor as well as a modification factor which increases or decreases the expenditures depending upon the activity (intense combat, normal combat, etc.) of the unit assigned a given weapon system. The result is ammunition expenditures per time period by individual DOD Ammunition Code. The following example will illustrate this computation:

Unit A has 100 weapon systems X
Rounds per weapon per day for weapon system X
is 10
Modification factors are as follows:

Intense Combat	2.0
Normal Combat	1.0
Reduced Combat	.6
Reserve Position	.2

5 days in one time period.

The computation of expenditures for this weapon system, assuming it expended one type of DODAC, would be as follows: (Assume intense combat) 100 weapon systems X 10 rounds per weapon per day X 2.0 (modification factor) X 5 days - 10,000 rounds expended per time period.

The highest expenditures of all types of ammunition as would be expected, occur during the early phases of a war with some sector expenditures for similar weapon systems significantly higher than another sector. Expenditures could be very high during this period for the howitzers, tanks, artillery, mortars and automatic weapons. The model output is identified by both DODAC expenditures and weapon system expenditures for all DODACs. This models' capability for such precise measurement makes it very critical to wartime analysis by being able to rapidly pinpoint not only when individual DODACs will become in short supply but the specific locations (ASPs, Depots, etc.) so that resupply shipments can be expeditiously made. Shortfalls in ammunition availability when compared with requirements can quickly be made.

3.8. Transportation Model

This model is one of the key elements in the entire SIGMALOG system. It computes requirements for transportation units to support the postulated operation, i.e., light truck companies, rail car trains, etc. and independent cargo carriers to receive, discharge, and move troop units, personnel replacements, and resupply material. This model has the capability to determine the location, time period of occurrence, and degree of deficiency of any transportation constraints. Roadways and railroads can be excluded from use for a given number of time periods to represent a more realistic situation as it would exist in wartime. The model is capable of accomplishing movements over five transportation modes--pipeline, air, rail, highway, and inland waterway. In accomplishing movements, the model computes the actual number of transportation units required and reports any deficiencies in net capacities. This permits a comparison between customer estimates and model-calculated requirements.

Movement between transportation nodes are designated as links. Each link is assigned a capacity in terms of short tons per day of movement. Actual movement by time period when compared to stated capacity identifies links

where deficiencies occur. Movements between transportation nodes are designated as either priority, preference or least-cost movements. The type of designation will either force a certain movement to occur before others or allow movements to occur until the capacity of links are exceeded. Capabilities of airfields and ports to accommodate influx of materiel can be examined in great depth. Penalties can be applied to ports, for example, to force the model to use certain ports first for certain type of cargo. Types of cargo to be identified could be troop units, replacements bulk pol, construction, etc. Port operations can be identified as to that cargo to be off-loaded in berths versus that cargo to be off-loaded in-stream or at anchorage. The number of port operations teams by time period can be determined at each port from the amount of cargo moved through the port. The same data can be obtained for airfields in comparing the amount of tactical air versus logistics air that can move through the airfield. Intra-theater air requirements can also be determined.

4. USES OF SIGMALOG

The general intent of the SIGMALOG System is to use it as an automated logistics planning and analysis tool providing various data that when properly analyzed would assist in making decisions concerning logistics support. One by-product of the application of the SIGMALOG System is the capability to create the nucleus of an automated data bank that can be used by all levels and branches of the Military.

Specific uses identified for the SIGMALOG Models are:

- 4.1. Validating mobilization plans.
- 4.2. Validating regional assignments.
- 4.3. Determining material support requirements by quantity (short tons), location, and time phasing.
- 4.4. Determining POW and Refugee flow patterns and planning for adequate class I, II and VIII support at collection points.
- 4.5. Determining hospital bed requirements with varying casualty rates.

- 4.6. Analyzing impact of personnel replacements on the transportation system.
- 4.7. Determining quantity and time period of CONUS resupply by class.
- 4.8. Determining critical losses for Major End Items.
- 4.9. Determining percentage and amount of air versus sea resupply for Major End Items.
- 4.10 Determining Ammunition requirements by time period, DODAC, weapon system, weapon system subset, region and Army levels.
- 4.11 Calculating expenditures of weapons systems at different firing rates and levels of activity.
- 4.12 Determining transportation units shortfalls in transportation networks requiring capacity overloads.
- 4.13 Determining optimum in place stock requirements by class.

This list could go on and on, but it should be evident by now that the potential to logistics planners is tremendous. Allowing for reprogramming of the models, many more applications can be accommodated.

On a larger scale the long-range goal is to use the SIGMALOG models to review and up date operational plans. It would be feasible to run a Command Post Exercise, Field Training Exercise, etc. and then test the results with a SIGMALOG analysis of the exercise to determine reliability of the exercise results.

METHODOLOGY FOR ASSESSING THE TRUE WORTH
OF PERFECT FORECASTS

Graham W. Winch,
Durham University Business School,
Mill Hill Lane,
Durham DH1 3LB, England.

The principal thrust in forecasting literature and research is towards the development of more accurate techniques. This effort results in practising managers being faced with an ever-expanding array of more complex methods. This contrasts with recent findings by the author that firstly, many managers regard the study of the role of forecasts in decision-making as of much more importance than technique development, and secondly, the study of the nature of the control processes is critical in the determination of required levels of accuracy.

This paper describes the use of the system analysis and simulation approach of System Dynamics in the examination of the effects of typical forecast error, especially systematic bias and random noise, on the behaviour and performance of three business control systems.

The conclusions drawn from this work are:

- (1) Many systems are likely to include mechanisms which result in high insensitivity to forecasting error, in such cases there is little utility for perfect forecasts and simple, cheap forecasting techniques are likely to prove adequate.
- (2) System Dynamics, used in this way, forms the basis for a methodology for the systematic evaluation of perfect forecasts, and the quantification of the effects of various types and levels of error on system performance.

Keywords: FORECASTING, EVALUATION, SYSTEM DYNAMICS,
METHODOLOGY.

1. INTRODUCTION

1.1 Purpose of this Paper

This paper describes the evolution of a methodology for assessing the true value of perfect forecasts in business control systems, and for determining the levels of forecasting accuracy required in particular situations to achieve satisfactory system control. The process of this evolution is firstly the analysis of current thinking concerning technique development and the utility of forecasts, secondly the description of an extensive program of simulation experiments designed to examine the effects of forecasting errors on system behaviour and control, and thirdly the formalisation of the methods used in the simulations into a rigorous and systematic methodology for evaluating the role of forecasts in any real business situation.

It is hoped that this methodology will go some way towards bridging the 'technology gap' between forecasting technique development and practical applicability; as Wood(1) has stated:

"Given the pace at which forecasting techniques have developed over the last few years, it is not surprising that at times the ability to generate forecasts seems to overwhelm any matching capability to evaluate them and integrate them into on-going decision situations".

1.2 Types of Forecast Errors and their Causes

Forecasting of all management processes is vulnerable to errors, these errors being functions of the natural variability of the real world or as a result of the forecasting processes adopted. It is clearly to reduce the size of these errors that such an array of complex forecasting techniques has been developed. It is not the primary purpose of this paper to concern itself with sources of forecast error but rather with their effects on system behaviour. It is appropriate here, however, to describe the nature of the various types of error that might be present in any forecast.

Bias is the systematic and consistent over- or under-estimation of a forecast variable. This error may arise from the character of the forecasting technique, e.g.

simple averaging and smoothing methods will lag behind a trend and unless corrected will under-estimate a rising trend and over-estimate a falling one. (This is known as "velocity error" in control engineering). Alternatively there may be some fault in the data collection procedure which introduces the bias, for example a sales revenue forecast might consistently fail to deduct agent's commission. A third possible source of bias is a function solely of the character of the individuals who prepare and/or use forecasts. They may be naturally optimistic or pessimistic and inadvertently allow this to be reflected in their forecasts, or there may be the deliberate adjustment of forecasts which produce bias, e.g. salesmen may under-estimate sales forecasts if they suspect that their figures will be returned to them as targets.

Noise or Random Error is the other major error component apart from bias, and all forecasts would be expected to include a component of this as residual error. This error arises from those aspects of the system, the forecast variable and the forecasting process, that occur on a purely random basis, and because they are random, individual values cannot, by definition, be predicted in advance. The best that can be hoped for is that a frequency distribution be identified which can indicate expected ranges and probabilities for various error levels. It might also be that where the residual error contains a component that is not strictly random - e.g. short term serially correlated errors - these might be identified. Again this error can arise from a number of sources - as a function of original time-series which depend on random factors like weather, and random errors introduced during data collection, forecast calculation or information transmission.

The Smoothness will also be a function of both the nature of the forecast process and the technique adopted. It would clearly be expected that aggregate forecasts would produce a smoother pattern than forecasts for individual elements, and similarly monthly figures would not be so sensitive to random day-to-day fluctuations as daily or weekly figures. Smoothing and averaging methods are probably the most common forecasting techniques and will clearly produce a smoother pattern than the original time-series. It may be desirable to smooth out day-to-day fluctuations with such a method, but the level of smoothing may be critical in terms of the technique's ability to deal with longer-term cycles or with changes in basic pattern like step rises.

The nature of information delay is obvious, but its cause may result in a constant delay e.g. from normal time involved in collection of data, calculation of forecast and transmission of results, or may have a random or variable delay, e.g. absence of forecasting staff, breakdown of telecommunication or computing equipment or inefficient use of information in decision-making.

1.3 Approaches to Forecasting in the Literature

In order to determine the balance of emphasis between technique development and the utility of forecasts and forecast methods, the author recently classified the articles on forecasting in issues of ten prominent management science and general management journals into five categories representing the differing approaches to forecasting adopted in each. The full analysis appears in Winch(2) but can be summarised as in Table 1.

Table 1 - Emphasis in Forecasting Literature

Classification of Article	Frequency
1. Mathematical Theory and Technique Development	19)
2. Comparison of Techniques in isolation using typical statistical measures of accuracy) 12)
3. Reports of the use of specific techniques in various applications) 16)
4. General philosophy (including state-of-the-art articles)	8
5. Articles recognising role of forecasting in control systems and implications for whole-system behaviour	9
	—
	64
	—

As can be seen, there is a very strong bias towards technique orientated papers with very little emphasis being placed on the relationship between forecasting and system behaviour. This clearly illustrates the problem identified in the quote from Wood in Section 1.1, and strongly contrasts with the views of practising managers as recorded in another survey reported in Winch(2). The overwhelming view of the managers was that, from their point of view, technique development has gone as far, if not further, than needed and that what is now required is research and development of an applied kind - case studies of successful applications of simple and complex methods, guidelines for the use of forecasts, and studies of the inter-relationships between the forecasting and decision-making processes.

One group of published work which does include consideration of this last aspect are reports of studies of business problems which have adopted the System Dynamics approach (see Section 2.2). These include the works of Barnett (3), Coyle(4) and Swanson(5) and a discussion of these together with an hypothesis as to why the System Dynamics approach tends to highlight such inter-relationships is given in Winch(6).

2. SIMULATION EXPERIMENTS WITH THREE BUSINESS SYSTEMS

2.1 Purpose of the Simulation Experiment Program

The program of experiments was designed to study the effects of errors or degradations, of the sort described in Section 1.2, in forecasts on the behaviour and performance of three business control systems. Simulation is an ideal medium for such a study, as it allows the researcher to isolate changes in behaviour that are due to particular changes in system structure or parameters without fear that other factors either internal or external will have changed.

It was also hoped that by adopting a simulation method which uses a structural approach, i.e. models causal relationships, it might be possible to relate particular aspects of the systems' responses to forecast errors, to structural features of the system, thereby gaining an understanding of the inter-relationships between the forecasting and decision-making mechanisms.

2.2 The System Dynamics Approach

System Dynamics was originally developed by Forrester and others at M.I.T. in the 1950's to study the dynamic (i.e. time related) behaviour of industrial systems, but has been used over the years to study all forms of 'managed' systems - socio-economic, ecological, social and judicial, besides business systems (see, for example, Forrester (7,8), Coyle(9), Sharp(10), Goodman(11) which give full discussions of the theory and a number of practical case-studies).

Most managed systems are highly complex structures of flows, physical and information, constraints, policies and decisions. System Dynamics is an approach for studying such systems through the examination and analysis of the feedback loops formed within the structure, and by relating this analysis to the behaviour of the systems both in reality and in simulation experiments. This process enables the analyst to consider the effects on system behaviour of changes in the basic structure, either caused by external influences or internal policy changes.

The basic promise of System Dynamics is that these managed systems, be they business or otherwise, can be considered analogous to physical systems like those, for example, in process plant. The analyst can therefore study their behaviour using the same theory and methods as the Control Engineer uses in his process control. In this respect he uses the concepts of flows and integration of flows into levels, of information feedback, the nature of positive and negative feedback loops, of control theory, and of conservative systems - i.e. that all materials have to be accounted for and cannot be lost or created in a process, only transformed.

The computer simulation of System Dynamics models can be carried out using any computer language which can accommodate the time-related difference equations but a number of purpose-written languages have been developed. The principal two are DYNAMO(12) the language originally developed at M.I.T. and DYSMAP(13), a compiler developed at the Bradford University System Dynamics Group (U.K.) which translates DYNAMO-type equations into FORTRAN, and runs in FORTRAN.

2.3 The Generation of Forecast Errors in the Models

The simulation models, as discussed earlier, are based on the actual structure of three business systems and naturally each includes forecasts of one or more variables (the models are described individually in 2.4). It is of no concern in this study how errors in these forecasts might arise, of interest is how the various types of error in each forecast affects the behaviour and performance of the systems.

The process of study is therefore to run the model with 'perfect' forecast(s) to obtain a base level of performance, and then to apply various degradation factors to each forecast to simulate different kinds of error - systematic bias, random error, information delay and level of smoothing. In this way changes in system performance can be related directly to particular kinds and magnitudes of error as the controlled conditions of the simulation ensure that all other factors and conditions remain unaltered.

For the purposes of this methodology forecasts must be divided into two categories - forecasts of exogeneous variables (i.e. variables which are fed into the model as time-series) and forecasts of variables generated internally. In the former case a 'perfect' forecast is simply obtained by using the same time-series as the actual input, with an appropriate lead-time representing forecast horizon; degraded forecasts being obtained by applying the appropriate error factor. With internal variables, the forecasts have to be generated mechanically for the 'perfect' case and congenital errors may be present - the degradations are therefore applied in addition to any already there. It is felt however that this limitation is acceptable for the comparative purposes of this work.

2.4 A Brief Description of Each Model

Clearly in a paper of this type there is no place for a lengthy discussion of the three models used. They are all described fully elsewhere(2). However, as each model was used to examine forecasting in a slightly different context it will be of value to highlight these differences by discussing each model briefly.

2.4.1 PRODINV Model

The first model examined represented a very simple production-inventory system for a one-product line with delivery from stock. The structure of the system is shown in Figure 1. This diagram maps out the inter-relationships between the variables in the system and is known as an Influence Diagram - see Coyle(4) for a full discussion of the mapping process. The principal control mechanism in this system is for production start rate and this depends on a forecast of demand for the product, and an error control mechanism based on the discrepancy between a desired inventory level and the actual level.

The main purpose of this simple hypothetical model was to study the changes in behaviour of the system produced by various errors in the demand/consumption forecast. To

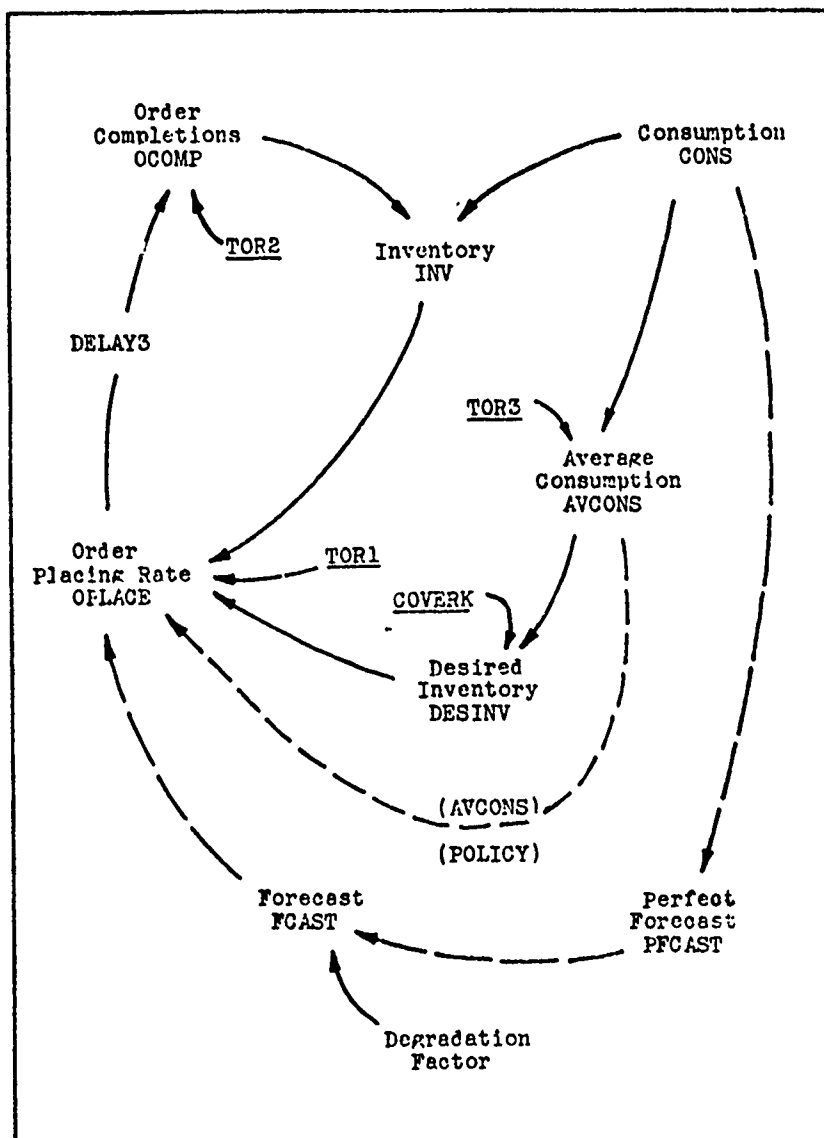


Figure 1 - Influence Diagram of Simple Production
Inventory System - PRODINV

this end the input demand patterns to the system were standard control engineering test patterns: a step (sudden change in level), a ramp (steady up- or downward trend), and a sine wave (representing cyclical pattern) and under examination were typical aspects of system transient and steady-state characteristics. The forecast degradations tested were smoothing, systematic bias and information delay.

2.4.2 TANKER Model

The second system examined concerned the chartering of crude-oil tankers. This highly complex model was originally developed by Coyle(4) for a major oil company, and has been adapted by the author to study the forecasts in particular. The system is concerned with the control of the company's tanker fleet in order to ensure the transportation of crude oil to refineries to satisfy fluctuating demand for oil products. Of critical importance is the composition of the fleet which consists of a small proportion of owned ships, a majority of time-chartered ships and a residue of spot-chartered tankers. A much simplified diagram showing the basic structure of this system is shown in Figure 2. The five underlined variables are the forecasts studied.

This system differs from the previous one in that it includes a direct quantifiable measure of system performance which can be used to compare the effects of error on the five forecast; the measure is CUMEXP - the company's total cumulative (over 10 years) expenditure on its tanker fleet. The exogeneous inputs to the system were the growth rate in GNP in North West Europe (the market for the oil-products) and the charter prices for spot- and time-chartering. These simulation experiments were concerned with the effects of bias and random error in each of the forecasts on system performance and factors representing these errors were introduced as discussed earlier.

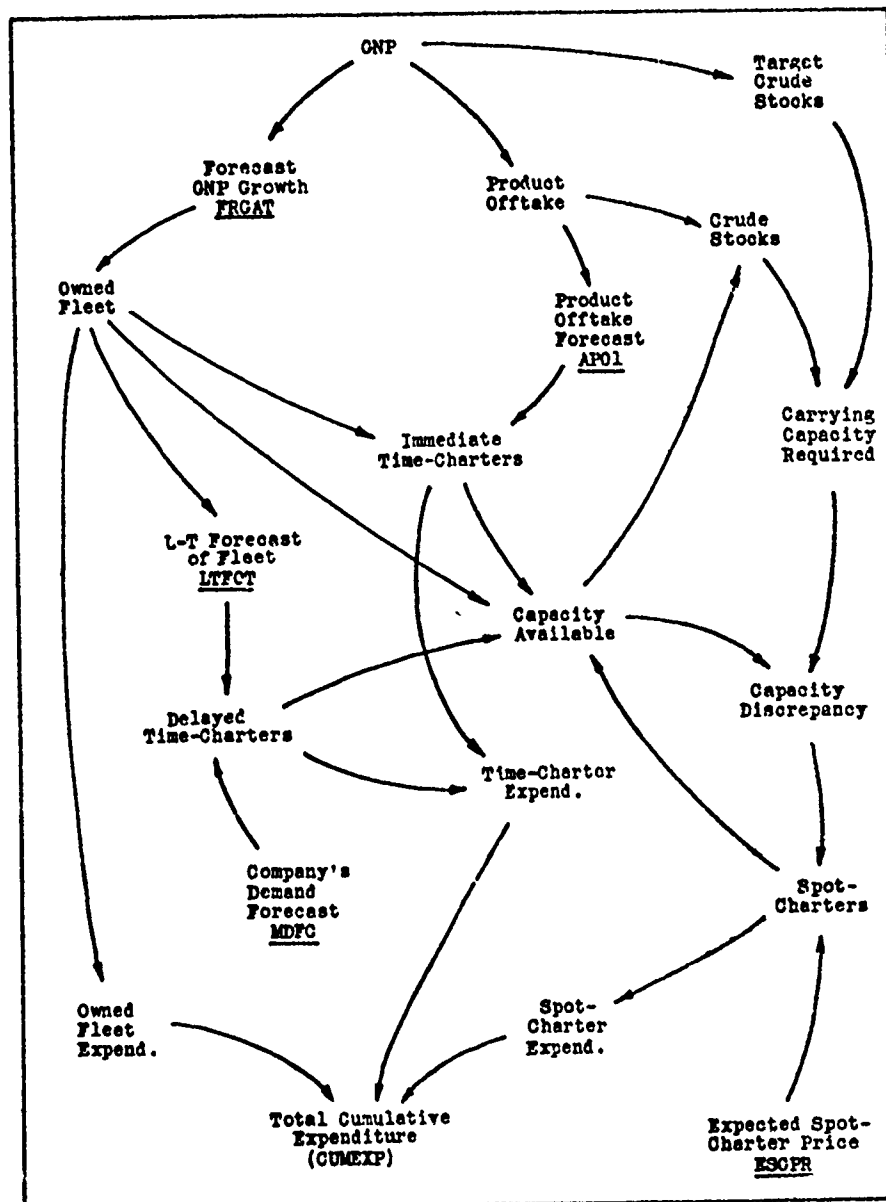


Figure 2 - Simplified Structure of Tanker Chartering System

2.4.3 BLDSOC Model

The final model under study was one developed by the author to examine the role forecasts could play in the control of the flow of funds from Building Societies. These societies are non-profit-making institutions concerned almost exclusively with providing finance for private house purchase for owner-occupiers and which provide nearly 80% of all such funds in the U.K. The cash-flow system is very simple. Societies can lend out funds they receive from investors (mostly small savers), repaid mortgages, and income from mortgage and investment interest. The assessment of performance of the system however is very complex - the societies must maintain certain accounting ratios, assess themselves in terms of growth in assets but probably most importantly by government and society, including their own investors, by their ability to supply a smooth and steady flow of funds. (The simplified structure of this system is also shown in Figure 3 with variables identified on Table 2).

This system differs again from the previous one in two ways. Firstly the complexity of assessment of performance means that no quantitative performance measures were really appropriate but rather subjective evaluation on the basis of graphical output was used. Secondly, the society modelled used current values rather than forecasts of sources of funds in its allocation procedures. This model was designed therefore to first assess which forecasts, if any, would prove of value in improving performance, and secondly the sensitivity of the system to errors in those forecasts.

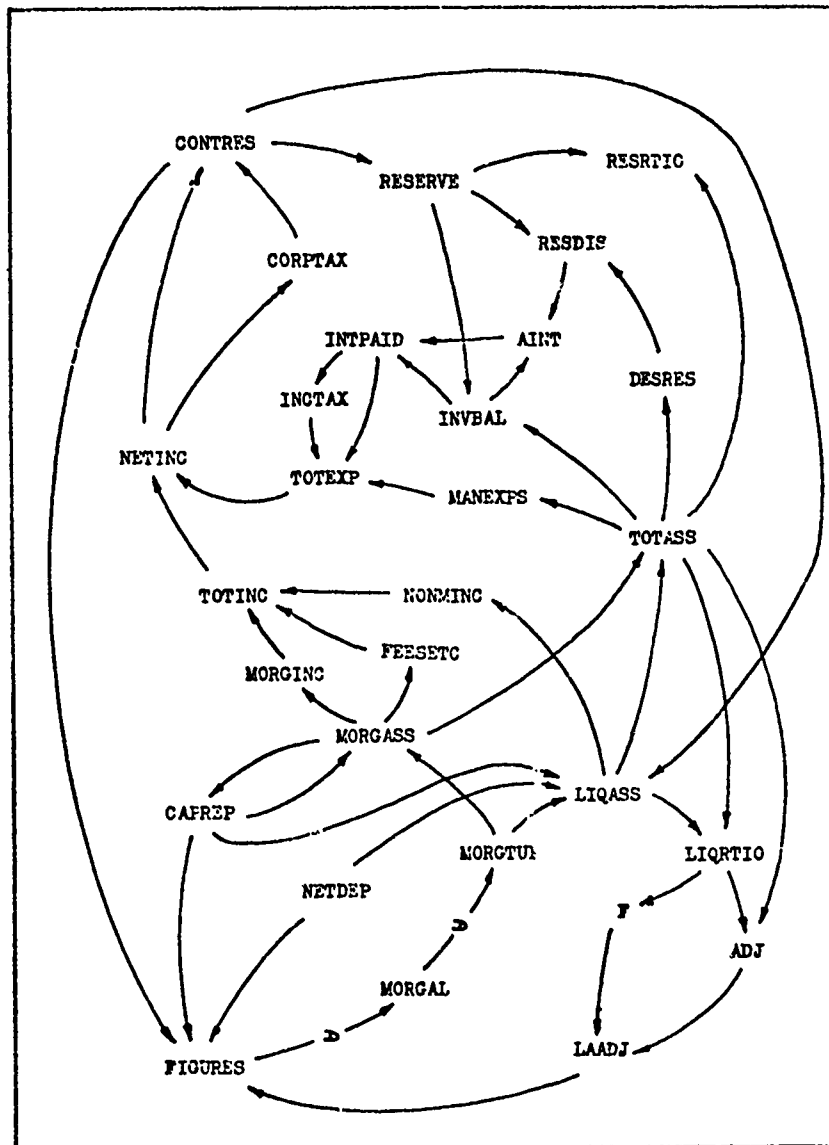


Figure 3 - Influence Diagram for Building Society
Model, BLDSOC. (Constants are not included)

ADJ	Adjustment required to control liquidity ratio	
AINI	Actual depositors' interest rate	
CAPREP	Repayment of loan capital	
COMINT	Commercial interest rate (C)	
COMPRAT	Composite income tax rate (C)	
CONTRES	Contribution to reserves	
CORPTAX	Corporation tax paid	
CTAX	Corporation tax rate (C)	
DEPTAB	Table of deposits	
DESLR	Desired liquidity ratio (C)	
DESRES	Desired reserves	
DESRR	Desired reserve ratio (C)	
EXPSFCT	Management expenses factor (C)	
F	Clipping function	
FEESETC	Income from fees, commission etc.	
FEEFCT	Fees income factor (C)	
FIGURES	Total figure-forms basis for mortgage allocations	
INCTAX	Income tax paid on behalf of depositors	
INTPAID	Interest paid to depositors	
INTRAT	Basic depositors' interest rate (C)	
INVBAL	Investors' (depositors') balances	
LAADJ	Adjustment in allocation to control liquidity	
LIQASS	Liquid assets	
LIQRTIO	Liquidity ratio	
LRADJT	Time for liquidity adjustment (C)	
MANEXPS	Management expenses (costs)	
MORGAL	Mortgage allocation rate	
MORGASS	Mortgage assets	
MORGINC	Income from mortgage interest	
MORGRAT	Mortgage interest rate (C)	
MORTUP	Mortgage take-up rate	
NETDEP	Net deposits	
NETINC	Net income (subject to Corporation Tax)	
NONMINC	Income from investment of liquid assets	
RESDIS	Discrepancy in reserves	
RESERVE	Reserves	Constants are indica- ted as (C)
RESRTIO	Reserve ratio	
TOTASS	Total assets	
TOTEXP	Total society expenses	
TOTINC	Total society income	
TUPTIME	Mortgage take-up delay (C)	

Table 2 - List of Variable Names for BLDSOC Model

3. DISCUSSION OF THE SIMULATION RESULTS

3.1 The Nature of the Results and Conclusions

It is a fact that over two hundred simulation runs were performed during the course of the simulation experiments and clearly it is neither possible nor desirable to discuss them all or even summarise them all at this time. It is the author's intention rather to discuss the major points and conclusions that emerge from the experiments and to illustrate these with examples where appropriate. A full and complete description of all the results appears in Winch(2), and the TANKER results are discussed in Winch(14).

The conclusions from the simulation program centre on three main points:

- (1) The general relationship between forecasting and forecast error and system stability.
- (2) The desirability of forecasting particular variables in attempting to improve system performance and general conclusions regarding sensitivity of systems to errors in forecasts.
- (3) The relationship between the response of systems to the forecasting functions and particular aspects of the system's structures.

3.2 Forecasting and System Stability

The experiments with the PRODINV model were basically concerned with the question of the relationship between forecasting, forecasting error and system behaviour and controllability, and accepting that the PRODINV model is of a very much simpler form than would normally be expected in business situations, some useful and extremely interesting conclusions can be drawn from this series of experiments.

- (i) The experiments established that the incorporation into the system of any stable forecasting systems based on consumption does not affect the basic stability of the system.
- (ii) Normal errors found in forecasts, viz bias, noise and information delay also do not affect the stability of the system. This implies that a system exhibiting unstable behaviour cannot be rectified by simply improving the forecasting function to remove these types of error.
- (iii) The using of a smooth forecast certainly does not necessarily mean smooth production.
- (iv) Throughout all these experiments there is absolutely no clear indication that the use of a 'perfect' forecast will lead to the best system performance (it has generally been assumed here that smoothness in production rates and the level of inventory staying close to the desired level have been the principal aims of the controller). In most cases there has been little or no difference between the result for any of the degraded forecasts against that for the 'perfect' case - on a number of occasions one or more of the degraded forecasts actually appeared to give better performance.

While not advocating the deliberate inclusion of errors in forecasts, it would seem to the author that in a case such as this, an attempt to eliminate any small errors that might be present in the forecast would be of somewhat doubtful benefit.

3.3 The Desirability of Particular Forecasts and System Sensitivity

The second two models studied - TANKER and BLDSOC - offered the opportunity for examining the value of particular forecasts in actual business systems, in the former case for existing forecasts in the latter for possible new forecasts.

Tables 3 and 4 summarise the results for the TANKER experiments and indicate the sensitivity of the system to random error and systematic bias in each of the five forecasts in turn.

Noise

In order to examine the system behaviour with respect to random noise in the forecasts, a noise function was applied, in turn, to each forecast. The noise function used was the DYNAMO NOISE function which produces random numbers uniformly distributed between positive and negative limits. The limits chosen for these experiments were +10%, +20%, +30%, +40%, +50% and ten simulation runs were conducted for each pair of limits. (The random numbers were serially correlated in so far as a SAMPLE function ensured the same error was used over monthly intervals).

Table 3 summarises the results for each of the five forecasts at each level of error. It is useful to compare the effects of various levels of error for each forecast with the base (no error) value for cumulative expenditure, and also to consider what sort of level of error would be acceptable for each of the forecasts.

Briefly one might conclude that for two of the forecasts, the performance is so unaffected that no serious concern need be taken at all, in two other cases errors of the order of +20% to 30% would be quite acceptable. Only in one case - the short-term forecast of product offtake - would any real attempt be needed to produce an accurate forecast.

Bias

The second set of experiments consisted of applying a number of bias factors to the same five forecasts. Again this work is not concerned with how the biases might arise, but is simply comparing system performance assuming them to be present. The bias factors applied were:

PERF: Bias factor = 1, hence the system assumes a perfect forecast is available

PESSI: Bias factor = 0.9, a pessimistic forecast assuming the forecast is always underestimated by 10%

Forecast Error	GNP Growth Rate	Med.- Term Demand Forecast	Long- term Owned Tonnage	Short- term Product Off- take	Short- term Spot- Charter Price
Base Value	45.43	45.43	45.43	45.43	45.43
10% Mean	45.50	45.50	45.47	45.61	45.50
S.D.	.02	.01	.06	.52	.01
20% Mean	45.51	45.73	45.45	45.97	45.52
S.D.	.03	.08	.15	.75	.01
30% Mean	45.50	46.08	45.63	45.50	45.52
S.D.	.04	.20	.31	1.28	.01
40% Mean	45.51	46.14	45.72	45.93	45.52
S.D.	.06	.37	.52	2.17	.01
50% Mean	45.53	46.66	45.85	48.04	45.53
S.D.	.08	.60	.55	2.03	.02

TABLE 3 CUMEXP (£100m) WITH NOISE IN FORECASTS

Forecast Bias	GNP Growth Rate	Mid.- Term Demand Forecast	Long- term Owned Tonnage	Short- term Product Off- take	Short- term Spot- Charter Price
PERF	45.43	<u>45.43</u>	45.43	45.43	<u>45.43</u>
PESSI	<u>45.26</u>	<u>45.43</u>	45.37	46.84	45.50
OPTIM	45.64	45.48	45.52	<u>44.44</u>	45.45
EXAGER	45.52	45.47	<u>45.25</u>	46.78	45.49
CONTRA	45.34	<u>45.43</u>	45.63	44.59	45.45

TABLE 4 CUMEXP (£100m) WITH BIASED FORECASTS

OPTIM: Bias factor = 1.1, an optimistic forecast
assuming the forecast is always
over-estimated by 10%

EXAGER: Bias factor varies, it is +10% when the market
is rising (i.e. when GNP growth is accelerating)
and -10% when falling.

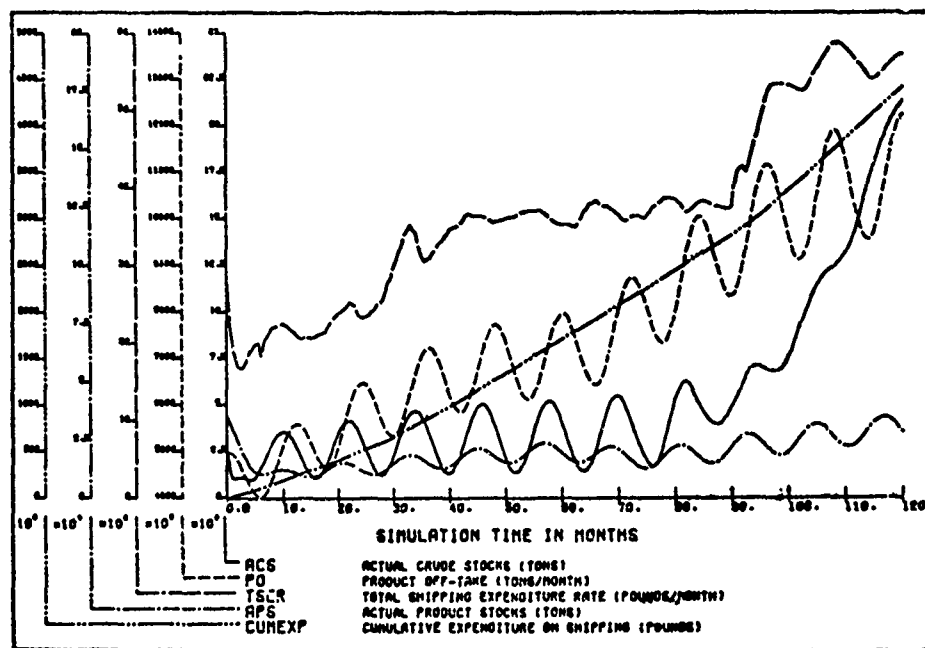
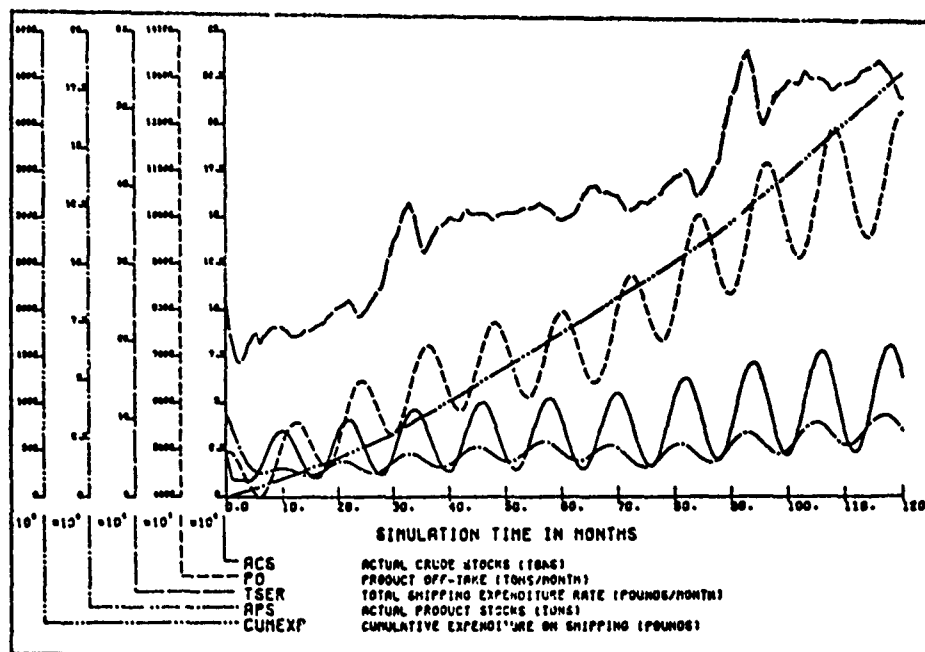
CONTRA: Bias factor varies, it is the reverse of the
EXAGER case and is -10% in a rising market and
V.V.

Again the results are summarised in Table 4, and show the simulated company's cumulative expenditure on shipping with each biased forecast, (the forecasts were biased one at a time) and the minimum cost result is underlined for each forecast. It can be seen from this table that none of the applied biases, again with the exception of the product offtake forecast, alters CUMEXP by more than about a half percent.

The author has also set out to discover the optimum values of forecast bias which produced minimum system cost, and the results of these experiments are discussed in Winch(15). Briefly they were:

Growth Rate in GNP	: FRGAT	: Optimum Bias = 0.032
Company Demand Forecast	: MDFC	: " " = 1.00
L/T Forecast of Owned Tonnage	: LTFOT	: " " = 1.59
Actual Product Offtake	: APOI	: " " = 1.06
Expected Spot-charter Price	: ESCP	: " " = 1.18

The implications of these results are quite startling - in one case there is a suggestion that a forecast should always be '0', and in another it even seems to suggest that the forecast should deliberately be biased up by 60% in order to achieve minimum cost performance! Figures 4 and 5 show the graphical output for the model with perfect forecasts and the optimally biased forecasts to enable comparison to be made between the general dynamic behaviour.



In the case of BLDSOC the study was two stage in that firstly the efficacy of possible new forecast(s) had first to be examined and, secondly, the sensitivity of the system to errors in these forecasts. As stated earlier assessment of performance was complex and basically subjective; however two means for quantitative assessment were devised. The first was the calculation of mean absolute deviation (MAD) and mean squared deviation (MSD) of the rate of new mortgage allocations about a central value - this gave a measure of smoothness of the flow of funds. The second produced a measure of the time that a critical accounting ratio lay outside particular limits (i.e. +1%, +2%, and +3% about the desired level of 15%) - this gave a measure of financial controllability.

Table 5 summarises the results for the first stage and it can be clearly seen that in only one case - FORDEP - would the inclusion of a forecast be expected to improve system performance. The improvement in smoothness of mortgage allocations (MORTGUP) and in control of the accounting ratio (LIQRTIO) can be confirmed by reference to Figures 6 and 7 which show the graphical output for the current no-forecast situation and where a perfect forecast of net deposits is available. (Net deposits - NETDEP - is the exogeneous input to this system).

With regard to the random error and bias that would be expected in any actual forecast of net deposits, the results were equally encouraging in that even with random error of up to +50% or with systematic bias of up to either +25% or -25% performance of the system was still significantly better than with the current practice of using current values only. This can be confirmed by reference to Figures 8 and 9 which show the graphical output for +50% Random Error (Noise) and -25% Bias.

The basic conclusions to be drawn from these sets of experiments are that systems are quite likely to exhibit insensitivity to forecast errors, and that those forecasts critical to system performance (either in existing or proposed forecasting functions) can be identified using System Dynamics simulations.

Table 5 - Performance of BLDSOC Model with
Prospective Forecasts Incorporated

Run Description	Deviations in MORTGUP		Proportion of run that LIQRTIO lay outside limits		
	MAD(10^6)	MSD(10^{12})	+1%	+2%	+3%
BASIC Current Practice - No Forecasts	2.74	12.96	74%	33%	15%
FORDEP Forecast of Net Deposits	1.94	5.43	40%	0	0
FORCAP Forecast of Capital Repay- ment Rate	2.77	13.17	61%	32%	8%
FORLIQR F/C of LIQ. RATIO Exp.Sm.t=6m.	2.41	9.24	51%	33%	22%
FORLIQR F/C of LIQ. RATIO Exp/Sm.t=12 m.	2.79	11.80	84%	55%	39%
FORLIQR F/C of LIQ. RATIO Trend Adj.Smooth	5.60	44.53	84%	70%	52%
FORDEP + FORLIQR (Combination)	2.04	5.93	41%	0	0

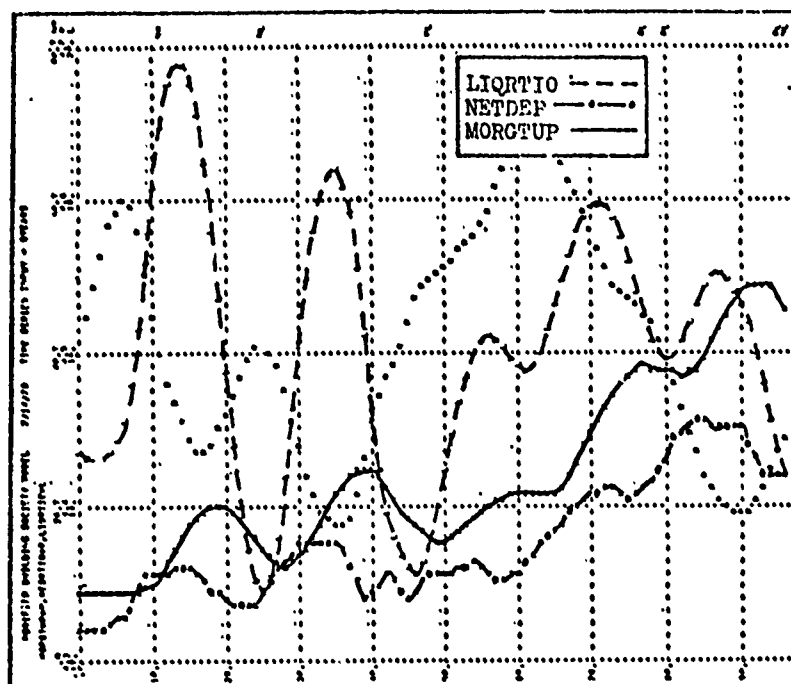


Figure 6 - Basic Run of BLDSOC Model, No Forecasts

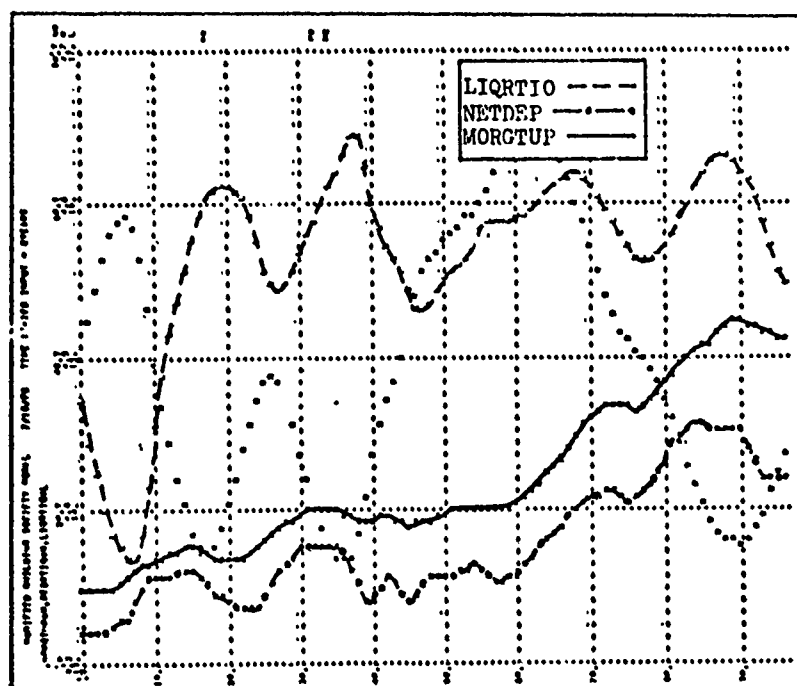


Figure 7 - BLDSOC with Forecast of Net Deposits

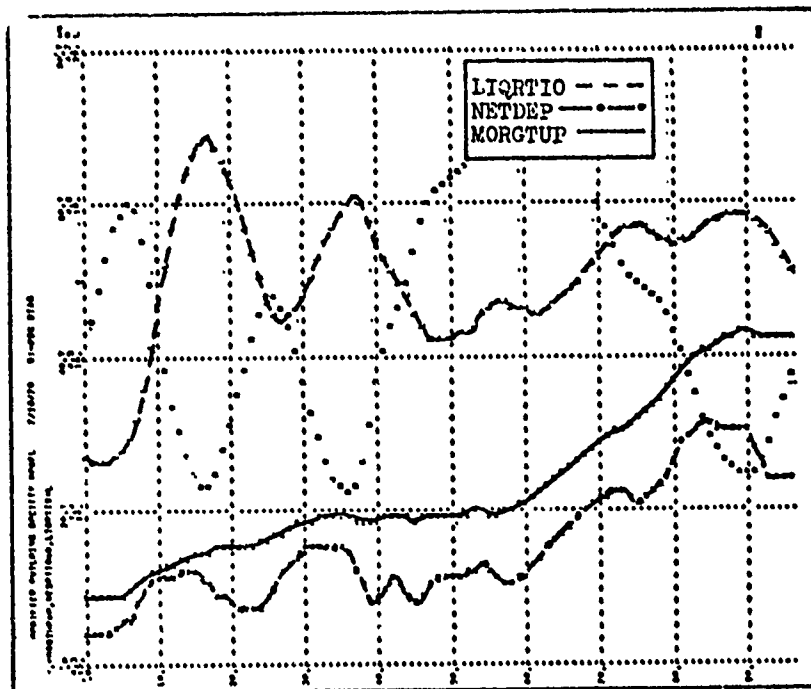


Figure 8 - BLDSOC with Biased Forecast (-25%) of Net Deposits

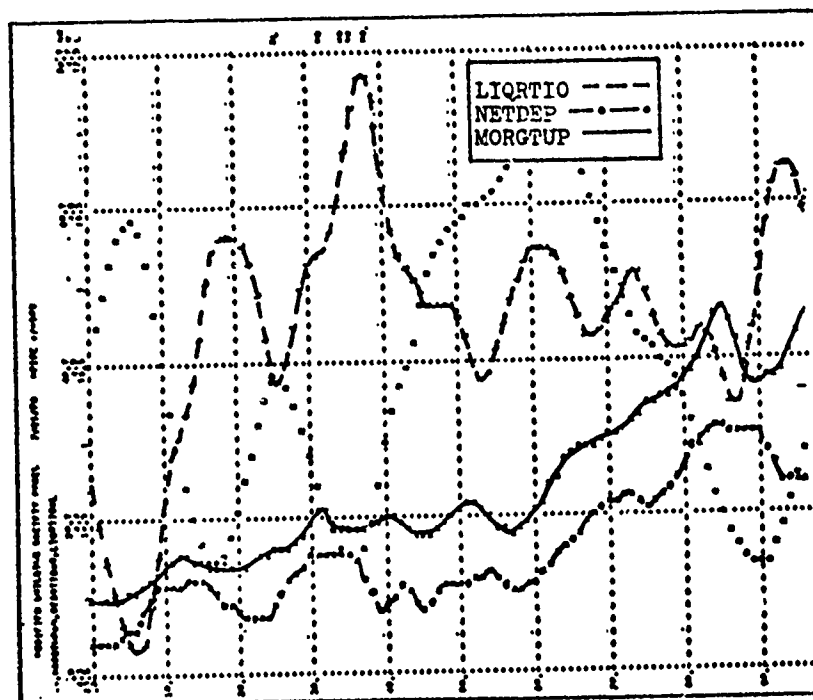


Figure 9 - BLDSOC with Noise ($\pm 50\%$) in Forecast of Net Deposits

3.4 Relationships between Response to Errors and System Structure

It was found during the original analysis of the results that the response of the systems to particular forecast errors frequently became less surprising or unexpected when a careful study of the system was made to establish how that forecast was used in decision processes or the nature of variables to be forecast.

Two examples illustrate this point well. In the first case one of the forecasts in TANKER was used to switch a policy from one option to another. With this mechanism it was only in rare marginal circumstances that an erroneous forecast would lead to selection of the 'wrong' option, and so over the ten-year run errors in the forecast produced only a very slight reduction in performance.

An example from the BLDSOC program concerns one of the prospective new forecasts. This forecast was of one source of funds available for allocation as mortgages, and it was a very smooth slow-changing flow. It would not therefore be expected that the forecast for 5 months ahead (the forecast horizon) would differ significantly from the current value, and so use of the current value would likely prove quite satisfactory in decision-making and little improvement would be expected from introducing a forecasting function.

In the studies the structural analysis was done as a 'post mortem' to identify the reasons why particular aspects of behaviour had occurred. It might be expected therefore that in further studies of this nature, many of the results might be predicted by such a prior analysis, thereby negating the need for some, if not all, of the simulation process. This analysis had certainly emphasised the validity of the opinions of the managers in the survey mentioned in Section 1.3 who stated that understanding of the role of forecasts in decision-making is probably more important than further technique development.

4. DEVELOPMENT OF AN ASSESSMENT METHODOLOGY

4.1 A Rigorous and Systematic Approach

It is a fact that the vast majority of books on forecasting and technique selection discuss desirable levels of accuracy and imply that some sort of cost-benefit analysis be carried out in order to find an optimum forecasting process which minimises the cost of inaccuracy plus the cost of forecast preparation. There is however precious little guidance let alone a rigorous method in any of these texts for determining costs of error or quantifying acceptable levels of error. The TANKER experiments described earlier enable the first to be done, though with the subjective assessment of BLDSOC it is less easy in that case. In both cases however the latter quantification is clearly possible.

It is therefore felt that a systematic, rigorous and realistic method of forecasting system assessment can be developed from the methods adopted in these cases for these reasons:

- (1) Forecasting is considered as an integral part of control processes.
- (2) The control systems are analysed to discover reasons for the effects of forecast error on dynamic behaviour.
- (3) Objective performance measures can be derived rather than subjective estimates.
- (4) Graphical representations of system behaviour can be produced in the simulation process to aid assessment or where quantitative measures are inappropriate.
- (5) 'Error' is broken down to enable separate consideration of the noise, bias, smoothing and information delay components.
- (6) The interactive (analyst/manager) process of S.D. model development aids credibility of conclusions.

4.2 Forecasting System Audit

On the evidence of this research it has been concluded that a meaningful evaluation of a forecasting system to ensure that optimum use is made of all available resources time, money, forecasting and other personnel, and techniques and methodologies - must enable answers to be found to the following six questions:

- (1) Does the inclusion of forecasts lead to improved system performance (measured in terms of explicit system objectives)?
- (2) Which particular forecasts do improve performance, and which make no difference? Do any forecast variables, or the use of a particular technique in certain conditions, reduce performance?
- (3) What level of accuracy (in terms of noise, bias, smoothing and information delay) can the system tolerate in the desirable forecasts before performance is likely to be significantly worsened? - This will of course form the basis for any cost of forecasting/value of accuracy considerations.
- (4) Are the desired levels of accuracy consistent with attainability in terms of (a) Techniques?
(b) Quality of Data?
(c) Stability of Environment?
- (5) What are the implications for performance of the use of attainable forecasts? (Costs and other constraints are of relevance here)
- (6) Can the system be modified to increase tolerance to expected errors or obviate need for particular forecasts?

(Note: As posed, these questions refer to the audit of an existing system, the questions and process would be identical, but for a change in tense, for a proposed system).

It is felt that the methods used in this research do constitute a systematic approach to providing the answers to these questions. The process, or Audit, is envisaged as comprising four stages:

- (1) Groundwork
- (2) Analysis of System Structure
- (3) Quantification of Effects through Simulation
- (4) Implementation and Evaluation of the Audit

(1) Groundwork

The first objective of this stage must be to enable a clear statement of system objectives to be made. This is the linchpin of the whole process for it is in terms of contribution to the achievement of these objectives that the forecasting system should be judged. Further the establishment of quantitative measures of performance related to the objectives is desirable for the objective analysis of results in Stage 3. It would also be desirable to obtain indications as to the expectations of forecasting held by individuals - this might serve to avoid some of the problems resulting from the dichotomy of forecaster / company objectives. A third area where information might be gained at this stage is regarding attainable levels of forecasting accuracy, as these are required in the answering of Question 4.

These areas of information are likely to be best studied through informal discussion plus the use of a questionnaire. In the latter case an element of the Delphi approach might be valuable in obtaining consensus opinion without the dangers normally associated with the behaviour of groups.

(2) Analysis of System Structure

A thorough knowledge of the control system in which the forecasts are used has long been regarded as an essential prerequisite for successful forecasting (see, for example, Chambers et al(16) pages 17 and 30-33). The purpose of such an analysis is in effect to inter-relate:

- (a) Objectives of the system
- (b) Nature of the forecast variables
- (c) The integration of the forecasts into the decision-making processes

The study of the feedback loops within the system is especially important in the analysis of system-wide implications of forecast incorporation and the identification of possible error amplifying or self-correcting mechanisms. The development of an influence diagram or similar representation of system structure is ideal for this analysis both in terms of facilitating the identification of important system components like feedback loops, and also in the communication with managers with direct responsibility and knowledge of the control system.

Much useful information about likely behaviour of the system in general, and particularly regarding forecasting implications, is possible from the influence diagram development and this initial qualitative analysis (as is often the case with System Dynamics studies). This stage will almost certainly make the major contribution to answering the first two questions of the audit, and may even indicate that quantification through simulation is not warranted. It might also be expected to give useful indications of expected results with respect to the other questions, particularly Question 6 which concerns possible modifications to the system to increase error tolerance.

This stage does not require the availability of a computer and suitable software, nor, except for the most sophisticated of loop analysis methods using control theory, does it demand any great expertise on the part of the analyst. (Although in fact the simulation process of system dynamics was also designed as a comparatively simple technique for use by the non-specialist). The linear mathematical analysis of the system is also likely to be of very limited value, not because it is inappropriate for complex business systems, but because of the general applicability of the conclusions regarding the relationship between forecasting and system stability developed by the author. It is further very difficult to perform rigorous analysis without a high level of expertise and familiarity in mathematical manipulation and control theory.

(3) Quantification of Effects through Simulation

The basic requirement of this stage of the Audit is to quantify and consolidate the indications and conclusions from the structural analysis. (This assumes that it was not decided after Stage 2 that this stage was unnecessary because of the nature of the system or not justified financially).

Such a quantitative analysis as this can provide empirical evidence as to the efficacy of particular forecasts, and also to the effects of forecast errors. Where the objectives of the system include a number of possibly conflicting components, performance indices may be derived with appropriate penalty functions and weights, as was done for the BLDSOC model. In such cases, and where quantitative measures are wholly inappropriate, the subjective evaluation of graphical output displaying the expected behaviour of key variables over time is likely to prove of most value. Although in such cases cost-of-error profiles cannot really be produced, nevertheless numerical estimates of error tolerance can still be made, on the basis of what is or is not judged as 'acceptable' behaviour on the basis of the graphical output (cf. the Building Society study).

Because the model used is based on the structure of the system, it means that in addition to the straightforward examination of forecasting with the existing control mechanisms, it is possible to consider forecasting where different mechanisms exist. This is likely to prove of value where changes in control policy are expected or as a direct contribution to answering Question 6 which is concerned with possible modifications to the system to increase error tolerance.

Besides the basic need for a computer, software and expertise, simulation is generally expensive. However in the System Dynamics process the most time consuming element is the development of the influence diagram which will already have been accomplished in Stage 2 of the Audit. This simulation stage is not likely therefore to be considered prohibitively costly. Another very important point is that a model once developed for the Forecasting Audit could be used and adapted for the study of other management problems.

(4) Implementation and Evaluation of the Audit

An audit such as the one proposed would certainly be costly, but the large amounts of resources currently employed in forecasting both by specialists and general managers means that careful planning and control of the forecasting function is vital. Cost estimates at this stage are very difficult as the modelling exercises described here were not designed as part of a Forecasting Audit but as either models developed originally to study some previous management problem or specifically developed to study forecasting but in a pure research context. As prerequisites for the modelling stage, a computer with appropriate software, plus documented procedures for the Audit must obviously be available. In terms of the modelling effort required, Coyle(4) estimated a total of 70 man-days was needed in the original development of the TANKER model, and the author estimates 30 man-days for BLDSOC or a model of similar complexity. Both these estimates include only the analyst's time and do not account for other managerial participation.

It is not envisaged that the process proposed is repeated at frequent intervals; regular evaluation of major forecasting efforts might be reasonable every five years, while clearly the Audit is appropriate when new major forecasting systems are proposed. There should of course be continuous monitoring for significant drifts in policies or constraints, and the implications in and from the forecasting function might be especially considered when it is known that major changes are planned in control policies. The maintenance of the model, perhaps through its use in studying other problems, would mean that regular evaluations could probably be performed at low marginal cost; further, continuing familiarity with the model on the part of forecasters and managers would aid the ongoing monitoring of the system.

As with most management science projects the utility of the Audit would be a function of the validity of the conclusions drawn and the efficiency of communication with managers both during model development and dissemination of results. The Influence Diagram has a particularly important role in communication for the circulation and continual reference to a well-drawn diagram will encourage both confidence in the model and the conclusions and can aid everyone's general understanding of the system (personal

experience of the author has confirmed the communication and comprehension value of influence diagrams in explaining an enterprise's operation).

Despite the rigorous nature of the Audit and its concentration on quantification, it is still likely to prove difficult to evaluate it itself. The principal problem would be in differentiating between the direct value in improved use of forecasts and indirect value in that the auditing process, particularly the system analysis and general system understanding, may itself lead to better decision-making. There is, of course, also no guarantee that the model developed is the best possible - validation can only be in terms of a subjective analysis of its ability to replicate the real system behaviour (perhaps by comparison using past time-series) and the acceptability of individual relations to those with responsibility and knowledge of the system.

A perpetual criticism of this type of simulation model is that although it purports to model dynamic behaviour it is static in that it is based on current system structure while real systems obviously develop over time, sometimes in unexpected directions. However the judicious use of time-varying or conditional relationships (e.g. with CLIP functions) should be able to mitigate the worst effects. Continuous monitoring especially by comparison with the Influence Diagram should ensure that significant and unexpected changes in the system structure are recorded at an early stage.

A task sequence chart is shown as Figure 10 with the individual tasks identified in Table 6. The purpose of this chart is to indicate how the various tasks within each of the four stages of the Audit described above are inter-related to enable the six Audit questions posed at the start of this subsection to be answered. Some tasks will of course contribute information relevant to more than one question, others provide general information or are simply pre-requisites for other tasks. Table 6 also summarises the principal contributions expected from the various tasks towards the answering of the six Audit questions.

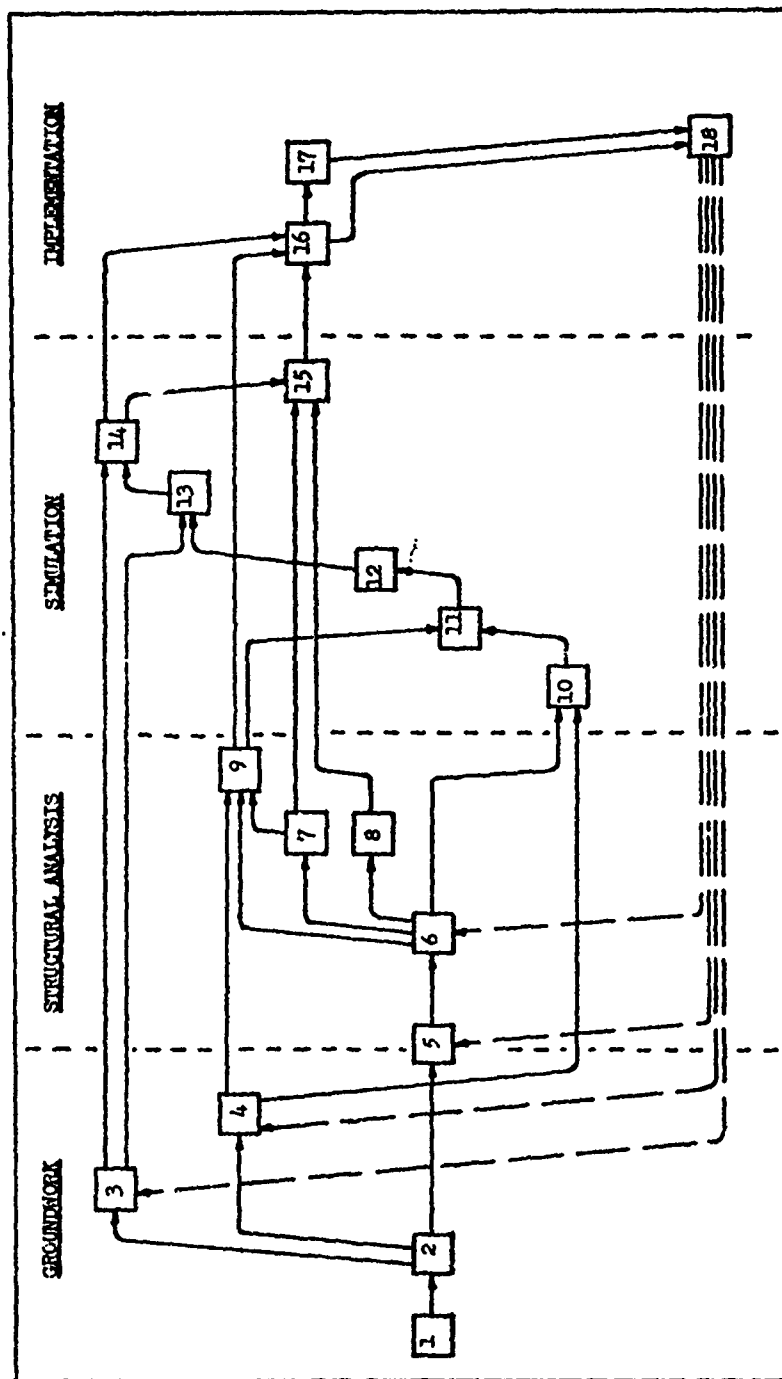


Figure 10 - Task Sequence Chart for the Forecasting System Audit
(tasks are identified in Table 5-1)

Task No.	Task	Audit Qu.
1	Preliminary discussion, examination of memos. etc.	
2	Detailed investigation of the control system, existing forecast practice, objectives, environment (use of questionnaire, Delphi-style if appropriate)	
3	Conclusions regarding attainable forecast accuracy through consideration of local resources (computer, software, personnel), data, environment	Q.4
4	Statement of company objectives	Q.1
5	Development of general understanding of the system	
6	Construction of Influence Diagram (including reference back to managers for approval)	
7	Identification of mechanisms likely cause system insensitivity to forecast errors	Q.1 & Q.6
8	Consider possible new mechanisms for increasing system insensitivity	Q.6
9	Preliminary (qualitative) conclusions as to value of forecasting in general and of particular variables	Q.1 & Q.2
10	Construct computer model on basis of ID and design experiments to examine value of forecasts	
11	Draw conclusions regarding value of forecasting and of particular variables in improving performance	Q.1 & Q.2
12	Develop cost-of-error profiles to examine system tolerance and acceptable errors	Q.3
13	Consider consistency of acceptable errors with attainable forecast accuracy	Q.4
14	Evaluate performance with attainable levels of forecast accuracy	Q.5
15	Consider improving existing mechanisms and introducing new mechanisms to increase insensitivity to error	Q.6
16	Implement findings with regard to: (1) Value of forecasting and of particular forecasts (2) Levels of acceptable forecast errors (3) Any necessary modifications to control systems	
17	Evaluation of Audit	
18	Monitoring of Control System, Forecasting System and the Environment	

Table 6 - Identification of the Tasks in the Forecasting System Audit, with an indication of each's contribution to the Audit Questions

5. CONCLUSIONS

Two major conclusions can be drawn on the basis of the work described in this paper. The first concerns the general implications of the simulation results and the second concerns the development of a formal forecasting system assessment methodology.

- (1) Many systems are likely to include mechanisms which result in high insensitivity to forecast errors. In such cases there is little utility for perfect forecasts and simple, cheap forecasting are likely to prove adequate. It is probable that such mechanisms could be identified by careful analysis of the system structure.
- (2) The process of System Dynamics study, with its emphasis on structure and causal relationships and comparatively simple simulation method, forms the basis for a methodology for the systematic assessment of forecasting functions. The Audit, as described in Section 4.2, can evaluate perfect forecasts hence indicating the value of particular forecasts in improving system performance, and can further quantify the effects of the various types and levels of error typically present.

6. REFERENCES

- (1) Wood, D. (1976) "The Quality of Forecasts - The Forecast/Decision Interface" CBR Conference - Forecasting for Business Decisions, Manchester Business School
- (2) Winch, G. W. (1978) "Forecasting and the Management of an Enterprise" Bradford University. Ph.D. Thesis, (submitted 1978)
- (3) Barnett, A. B. (1973) "A System Dynamics Model of an Oilfields Development", Bradford University, Ph.D. Thesis.
- (4) Coyle, R. G. (1975) "A Systems View of Forecasting" in Practical Aspects of Forecasting. Proceedings of 1973 RSS/ORS Conference
- (5) Swanson, C.V. (1971) "Evaluating the Quality of Management Information" WP 538-71 Sloan School of Management, M.I.T.
- (6) Winch, G. W. (1975) "The Dynamic Implications of Forecasts", DYNAMICA 2, 1
- (7) Forrester, J.W. (1961) "Industrial Dynamics" M.I.T. Press
- (8) Forrester, J.W. (1972) "Principles of Systems" Wright-Allen
- (9) Coyle, R. G. (1977) "Management System Dynamics" Wiley
- (10) Sharp, J. A. (1974) "A Study of Some Problems of S.D. Methodology", Bradford University, Ph.D. Thesis.

- (11) Goodman, M. R. (1974) "Study Notes in System Dynamics", Wright-Allen
- (12) Pugh, A. L. (1970) "DYNAMO II Users' Manual", M.I.T. Press (DYNAMO III now available)
- (13) Ratnatunga & Stewart (1977) "DYSMAP Users' Manual" Bradford University
- (14) Winch, G. W. (1977) "Feedforward Analysis in the Assessment of System Performance", TIMS/ORSA Meeting, Atlanta, Georgia
- (15) Winch, G. W. (1976) "Optimisation Experiments with Forecast Bias" DYNAMICA 2,3
- (16) Chambers, Mullick et al (1974) "An Executive's Guide to Forecasting", Wiley-Interscience

A Study on Performance Evaluation
for A Computer System Through Simulation

Louis K. Chow and C. Y. Chuang^{*}
Graduate School of Information Engineering
Tamkang College, Taiwan
Republic of China

ABSTRACT

In this paper, the resource allocation of an IBM 370/135 computer system is analyzed. The analysis is carried out by using queue theory to set up system model and by employing GPSS in computer simulation. Several situations of different resource allocations are then evaluated.

I. INTRODUCTION

Recently, the evaluation of computer performance is a topic of resurgent interest both within and outside the computer community. Especially, users of computer systems are eager to explore and to quantify the performance of machines they install and program.

* C. Y. Chuang is now with DP Center, Chinese Petroleum Corp., Taipei, Taiwan, Republic of China

It is a modern computer system that the computer and its linked peripheral equipments are organized to keep resources highly and efficiently utilized by implementing the multiprogramming mode through the control of a sophisticated software operating system. The hardware of a computer system is generally composed of a central processing unit (CPU), a primary memory, and some input/output facilities such as magnetic disk, tape, printer, and card-reader etc. The configuration is different for each installation because of different environment and applications. In this paper, an IBM 370/135 system is studied, which was installed in Tamkang College of the Republic of China during 1973-1977 period. The performance analysis is carried out by using queue theory to set up system model and by employing GPSS (General Purpose Simulation System) as a programming language in computer simulation. Through changing some system characteristics in simulation such as number of memory partitions, channels, and peripheral equipments, the variations of resources utilization and bottlenecks of queue are then observed. The results definitely provide the information for decision-maker to effectively allocate the hardware resources and thereby increase the service-ability of the system.

II. DESCRIPTION OF SYSTEM

The early computer systems had three basic components: (1) the CPU (Central Processing Unit); (2) the main storage unit (memory); and (3) I/O devices (peripheral equipments). The components, for instance, are interconnected as illustrated in Fig. 1.

As computer systems evolved, performance was upgraded by increasing CPU speed and memory size. For example, memories with 1,000 bytes at one msec per access have been replaced by models with over a million bytes at less than one usec per access. In similar fashion, CPUs have improved so that the

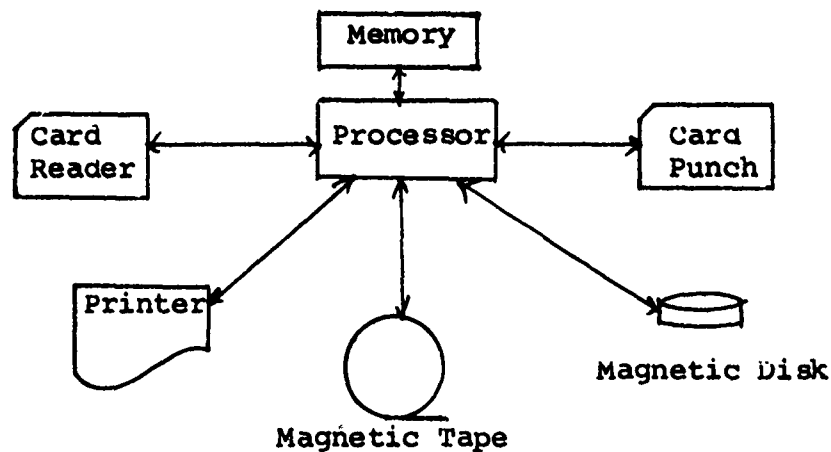


Fig. 1 A Simple Computer System

typical instruction time is less than one usec. However, the peripheral equipments, due to their electromechanical operations, are not able to improve their processing speed competitive with CPUs and memory elements. The disparity in speeds between the I/O devices and the CPU motivated the development of I/O processors (also called I/O channels since they provide a path for the data to flow between I/O devices and the main memory). I/O channels are specialized processing units intended to operate the I/O devices. Since these units may be simple, specialized, and not too fast, they are generally much less expensive than a conventional CPU. If all input and output are executed via the channel, the CPU is free to perform its high-speed computations without wasting time on slow I/O operations such as reading cards. Furthermore, since it is possible to be operating several channels simultaneously, many card readers, punches, and printers may function at the same time.

Although the basic idea remains the same, I/O channels come in all shapes and sizes, ranging from very simple processors to highly complex CPUs. At this time, there are two types of

channels in use. That is selector and multiplexor channels.

A selector channel can serve only one of its devices at a time. These channels are normally used for very high-speed I/O devices, such as magnetic tapes and disks. Thus each I/O request is usually completed quickly and then another device selected for I/O. But, a multiplexor channel is an I/O processor which can serve many devices simultaneously. It is able to accomplish this only for slow I/O devices, such as card readers and printers.

In recent years, the modern computers have employed multiprogramming mode in their operating systems' (OS) design for the purpose of supervising the operations mentioned above. The multi-programming is a technique that allows the concurrent execution of more than one program in a single computer system. That is, it allows the I/O operations of one program to be overlapped by the processing of other programs. That is, when a program has to wait for the completion of an I/O operation, the operating systems sets the program in the wait state and selects another program for execution on the basis of its priority and readiness to run. Thus, multiprogramming mode balances the difference between the speed of the central processing unit and the relatively slower speed of I/O devices, and thereby improves the overall throughput of the system.

Efficient use of the system relates not only to the degree of CPU activity but also to storage management. During system generation, storage may be allocated to partitions to accommodate the programs that will be executed in them. At times, only a portion of the partition is used by the program executed. DOS/VS of IBM's machine can automatically balance the storage demands made by programs by making processor storage not being used by one program available to a program in another partition as required.

DOS/VS can support up to five separate partitions[1] in each of which a problem program can be executed. Thus, up to five problem programs can be executed concurrently within the system, in which each program gets the priority associated with the partition in which it is executed.

The model system employed in this study is an IBM 370/135 computer which was installed in Tamkang College of the Republic of China during 1973-77 period. This system had equipped with POWER/VS (Priority Output Writers, Execution Processors, and input Readers/Virtual Storage) which is a spooling(simultaneous peripheral operations on line) program that also includes job scheduling. The POWER/VS program performs spooling of unit record data in DOS/VS which can reduce CPU dependency on mechanical equipment by using faster disk devices or magnetic tape units as intermediate storage. Then during execution of the problem program, data is read from and written to intermediate storage; program execution does not have to wait for unit I/O operations. After execution, printing and punching are done while other problem programs are executing.

The system organization of this model system is illustrated in Fig. 2. A brief description of its components is tabulated in Table A.

III. MODELLING AND SIMULATION

The data-processing operations of the model computer depicted in Fig. 2 which was equipped with particular features of job scheduling and memory management can be realized as a queue system of single server with $N-1$ spaces[2]. In this queue system, the arriving customers are computer programs which line up in channels waiting for CPU service; the server is CPU and main storage; and the queues are formed up in all channels or I/O processors.

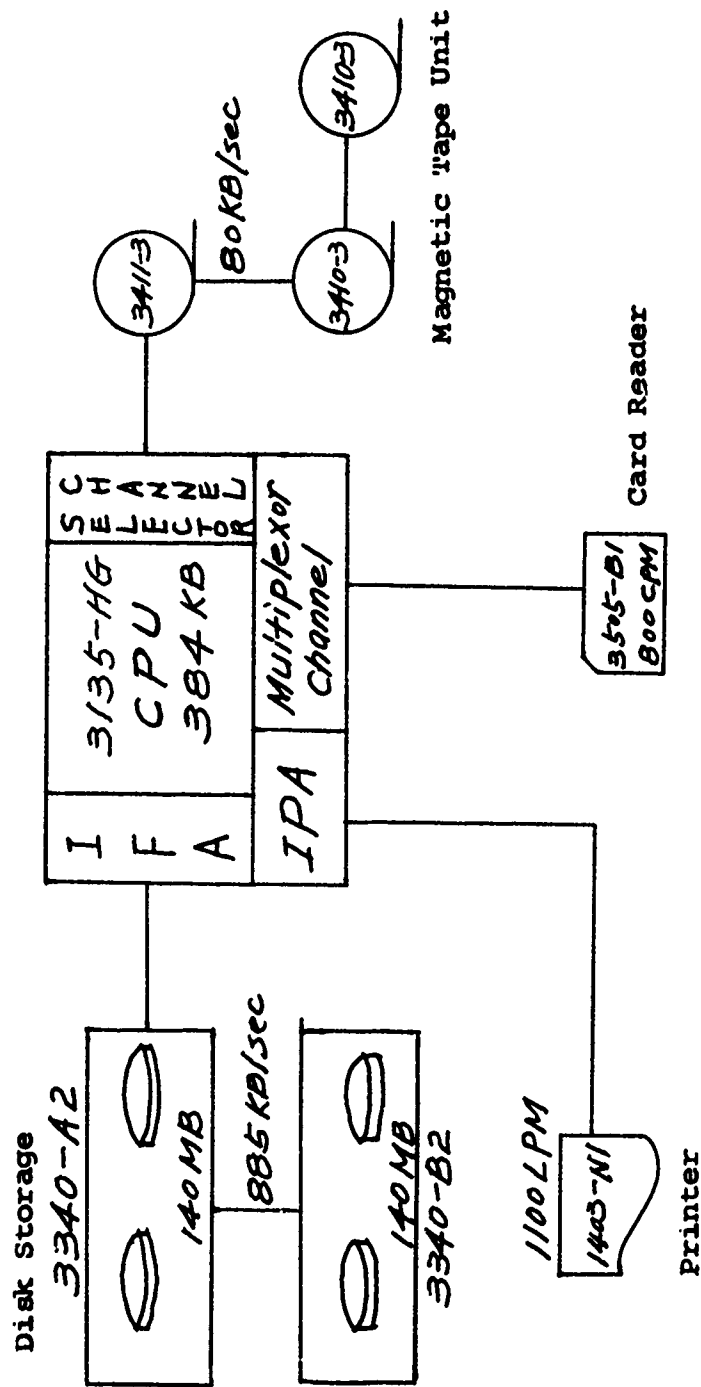


Fig. 2 The configuration of Model Computer

Table A

element	description
3135-HG	Central Processing Unit Capacity: 384K bytes
3046-1	Power Unit
3505-B	Card Reader Speed: 800 cards per minute
1403-N1	Printer Speed: 1100 lines per minute
3340-A2	Disk Storage, 2 Drive & Control Capacity: 140 Million bytes (MB) Data rate: 885K bytes per second (KB/sec)
3340-B2	Disk Storage, 2 Drives Capacity: 140 MB Data rate: 885 KB/sec
3411-3	Tape Unit & Control Data rate: 80 KB/sec
3410-3	Tape Unit, 2 Sets Data rate: 80 KB/sec
IFA	Integrated File Adapter, an I/O channel for disk units.
IPA	Integrated Printer Adapter, an I/O channel for printer.

In order to determine the statistical distribution of CPU service times, the execution-time intervals of 520 job steps were taken from system accounting log. It was proved by testing for goodness of fit that the CPU service times of this model system is exponentially distributed with mean equal to 15.68 msec. The histogram was plotted by Wang/2200 and is shown in Fig. 3. In the mean time, the frequency of I/O inquiry in each channel is also analyzed and listed in Table B.

In order to complete the computer simulation, the access times in each channel should also be given. Since there are 8 K bytes in each track

START I/O DISTRIBUTION

```

0 I*
1 I*****
2 I*****
3 I*****
4 I*****
5 I*****
6 I*****
7 I*****
8 I*****
9 I*****
10 I*****
11 I****
12 I***
13 I**
14 I***
15 I*****
16 I*
17 I*
18 I*
19 I**
20 I
21 I
22 I*
23 I*
24 I**
25 I*

```

X-AXIS : TIME (UNIT=1/300 SEC.) Y-AXIS : FREQUENCY

MEAN= 4.796797512633

ST. DEVIATION= 4.3874405219

Fig 3 Statistical Histogram of Service Time

Table B

Channel	IFA	Selector Channel	IPA	Multiplexer Channel
relative frequency	40%	3%	47%	10%

of 3340 disk drive, the largest size of each block buffer which accesses to CPU is also 8 K bytes. Each block size is assumed to be utilized equally likely and is similarly assumed for 3410 tape drive. Therefore, the access times of each I/O operations through IFA and selector channel can be obtained and are shown in Table C.

Table C

Block size	1KB	2KB	3KB	4KB	5KB	6KB	7KB	8KB
Relative frequency	12.5%	12.5%	12.5%	12.5%	12.5%	12.5%	12.5%	12.5%
Access time of IFA (MS)	2.4	4.8	7.2	9.6	12	14.4	16.8	19.2
Access time of selector channel (MS)	24.6	37.2	49.7	62.2	74.8	87.3	99.9	112.4

For IPA and multiplexer channel, the approximate numbers of cards and lines per block are 10, 6, respectively [1] which are equivalent to 750 msec and 330 msec as listed in Table D.

The exponential distribution of service

Table D

Channel	IPA	Multiplexer Channel
Spooling time (MS)	330	750

times can be put into such a form that, given a value from a 0-1 uniform distribution which can be called out in almost every medium size computers, the corresponding interarrival times can be directly computed. The pertinent equation is [3]

$$IAT_{\text{sample}} = (IAT_{\text{avg}}) [-\log_e (1 - RN_j)] \quad (A)$$

where IAT_{sample} stands for the sampled inter-arrival-time value; IAT_{avg} is the average inter-arrival time in effect; RN_j is the name of one of the GPSS uniform random-number generators, where the choice of j , as usual, is up to the analyst.

The actual simulation was run in IBM 370/135 system by employing the special-purpose language GPSS(General Purpose Simulation System). Making use of equation (A) and using the data tabulated in Tables B-D, the simulation was executed by dealing the number of partitions as a changing parameter. The results are shown in Table E, in which the average utilization, service time, and waiting time of each key component are tabulated. As the number of partitions changes from one to four, we can see that utilization, service time, and waiting time of CPU appears remarkable increase. Especially, when there are four partitions the CPU becomes heavily utilized and obviously the jobs will pack up seriously for waiting for service. The reason is that when there are three partitions the IFA channels have already reached almost 70%

Table E

component	average utilization				average service time (msec)				average waiting time (msec)			
	A	B	C	D	A	B	C	D	A	B	C	D
CPU	0.348	0.347	0.459	0.997	15.582	18.975	19.082	34918.977	0	0.507	1.714	14348.68
IFA(1)	0.059	—	—	—	9.382	—	—	—	0	—	—	—
IFA(2)	—	0.531	0.694	0.003	—	129.468	129.749	395.166	0	0	26.546	43.500
Selector Channel	0.031	1.033	0.042	—	63.324	63.887	64.294	—	0	0.149	0.542	—
IPA	0.517	0.434	0.576	0.002	69.388	329.886	329.696	330.000	0	34.472	60.790	132.666
Multiplexor Channel	0.443	0.129	0.170	0.002	88.812	750.000	750.000	750.000	0	36.161	64.219	0

A: mono-programming

B: two-partitioned multiprogramming

C: three-partitioned multiprogramming

D: four-partitioned multiprogramming

of utilization (Table F1) which can make the system unstable [4]. So, we tried to add two more IFA channels into the three-partitioned system and look the situation again. It shows (Table F2) as expected that the utilization of IFA only has 0.376 which is about one-half of its original quantity. However, this action also makes the CPU slightly increase its utilization. Furthermore, for reducing the waiting time of the IPA and multiplexor channel, the actions of respectively adding one more IPA and multiplexor channel were adopted and simulated. The results are shown in Table F3 and F4, respectively.

From the above analysis it appears that the optimal configuration for the investigated IBM/135 system are three partitions, four IFAs, one selector channel, IFAs, and two multiplexor channels.

IV. CONCLUSION

Computer simulation is a worthy technique for studying the optimal allocation of the precious resources. In particular, the resources of computer systems, due to the high prices and increasingly popularity, are urgently needed to give notices on the performance evaluations in order to increase the cost-effectiveness. Although the performance depends on the characteristics of each computer system and the environment where it is installed, it is hoped that this paper could be served as an example for the same kind of study.

V. REFERENCES

1. "DOS/VS System Management Guide", GC33-5371-4, File No. S370-34, IBM.
2. S. K. Gupta and J. M. Cozzolino, "Fundamentals of Operations Research for Management", Holden-Day, Inc., 1975.

3. T. J. Schriber, "Simulation Using GPSS", 1974.
4. "IBM S/370 Model 135 Channel Characteristics", IBM.

Table F1

Component	Average Utilization	Average Service Time (msec)	Average Waiting Time (msec)
CPU	0.459	19.082	1.714
IFA(2)	0.694	129.749	26.546
Selector Channel	0.042	64.294	0.542
IPA	0.576	329.696	60.790
Multi-plexor Channel	0.170	750.000	64.219

Table F2

Component	Average Utilization	Average Service Time (msec)	Average Waiting Time (msec)
CPU	0.499	19.060	1.278
IPA(4)	0.376	128.217	0.0
Selector Channel	0.041	05.219	1.149
IPA	0.623	329.766	78.698
Multi-plexor Channel	0.191	750.000	66.456

Table F3

Component	Average Utilization	Average Service Time (msec)	Average Waiting Time (msec)
CPU	0.519	18.683	1.086
IFA(4)	0.383	121.446	0.0
Selector Channel	0.048	61.566	1.014
IPA(2)	0.328	329.450	2.963
Multi-plexor Channel	0.181	750.000	59.954

Table F4

Component	Average Utilization	Average Service Time (msec)	Average Waiting Time (msec)
CPU	0.514	18.951	1.943
IFA(4)	0.391	127.841	0.0
Selector	0.047	65.175	0.767
Channel			
IPA(2)	0.319	329.511	3.054
Multi-			
plexor	0.093	750.000	0.0
Channel(2)			

SMART - Scientific Management Analysis and Review

Techniques for Local Financial Institutions -

MASAYUKI AKIYAMA

**System Laboratory of Information Processing
Systems Development Department**

FUJITSU LIMITED

1-17-25, Shinkamata, Ohta-ku, Tokyo, Japan

ABSTRACT. SMART is an application program package which is jointly developed by Fujitsu and Japan's eleven (11) local financial institutions for the purpose of supporting management's decision making.

This package consists of the following four applications;

Deposit Yield Forecasting Model (DEPY)

Loan Yield Forecasting Model (LONY)

Budget Planning Model (BUDGET/P)

Budgetary Control Model (BUDGET/C)

The sphere, and system specifications of SMART are discussed in this paper.

1. INTRODUCTION

The EDPS in Japan's financial institutions has achieved a remarkable progress from a primitive batch processing to an online real-time system.

The computer systems have become indispensable for operational (daily) transaction processings. Also, a new computer usage, such as the computer application for management decision making, has been tried in various financial institutions.

Business conditions in banking have become worse due to the decline of Japan's economic growth and the decrease in the growth rate concerning the amount of deposits or loans in many financial institutions.

The development of the computer-assisted management system is one of the most important task for systems engineers in their corporations. Accordingly, SMART was developed as a software tool for this purpose. However, the difficulty in the development of SMART is that the central idea of system design should be specified based upon the managers' thinking process.

To overcome the above difficulty, a project team was organized in 1974 by Fujitsu and eleven (11) financial institutions for the development of SMART. Not only the systems engineers but also the end users (managers) joined the team and determined the system requirements.

This paper presents an overview of SMART in the following section. The management framework is described to make the scope of SMART clear. Moreover, Section 3. describes the system functions of SMART which consist of four applications.

2. MANAGEMENT FRAMEWORK

In the early stage of SMART development, we had to design the management framework in order to determine the systems requirements.

The framework shown in figure 1 consists of the following three parts;

- Management Functions
- Classification of Management Problems
- Decision Making Procedure

The management functions, in other words, the role of managers in their organization, are considered as a management process, i.e. planning, control, and evaluation.

The management problems can be classified into two groups: one is strategic decision making problems and the other tactical problems. The difference between the two is that the former concerns the application of the management resources, and the latter the allocation of the given resources.

The decision making process is usually classified into three stages, i.e. recognition, alternatives-search and alternatives-selection.

Figure 2 shows an actual management framework in the financial institutions, where only two parts are selected representing the above-mentioned three parts. In order to determine the strategic or tactical policy, for example, it is represented in Figure 2 that the governmental regulations, trends of national/regional economy, or competitors business should be analyzed.

Figure 2 also illustrates the scope of SMART. As is shown, SMART deals with tactical problems from the viewpoint of the "Classification of Management Problems".

The managers and their staff can obtain valid informations to solve the tactical problems through SMART for the planning and evaluation processes.

SMART can also support the decision makers concerning the "recognition" and "alternatives-selection" stages. However, the "alternatives-search" is almost left to the decision makers own capabilities. In this sense, SMART can be considered as a man-machine interaction system for

management decision makings.

The following section describes in details the system functions of SMART.

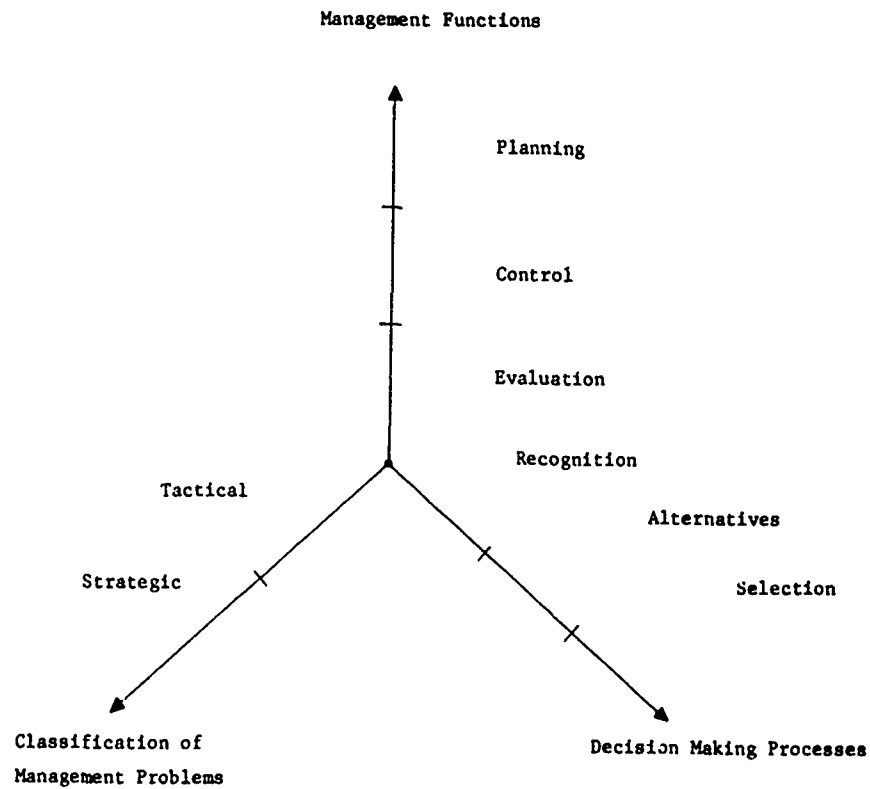


Figure 1 Framework

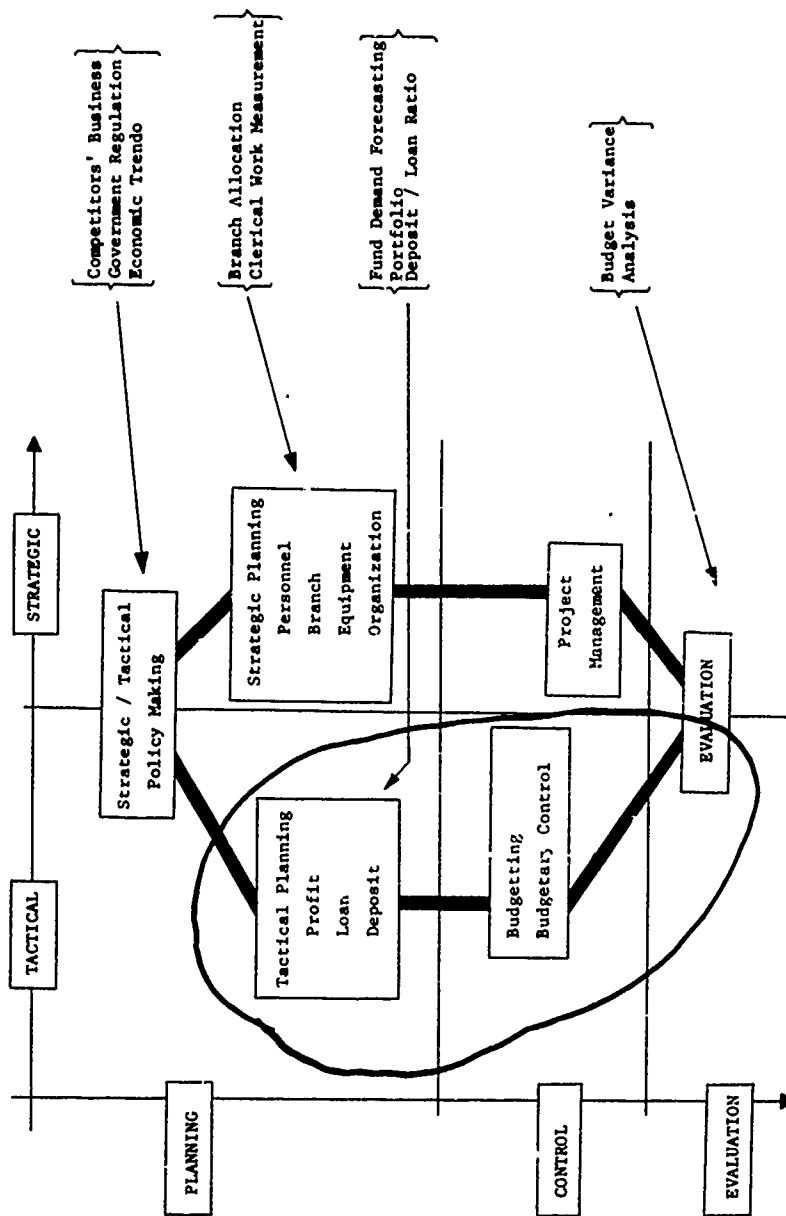


Figure 2 Framework for Financial Institution

3. SYSTEM FUNCTIONS OF SMART

3.1 Outline

SMART consists of four applications as illustrated in Figure 3.

Deposit Yield Forecasting Model (DEPY) and Loan Yield Forecasting Model (LONY) are support tools for the "alternatives-search" stage in the decision making process for business planning. These applications calculate the estimated yield based upon the current assets or funds rate-term mix. Budget Planning Model (BUDGET/P) and Budgetary Control Model (BUDGET/C) are financial simulation models, which were developed to test the alternatives in the final decision making process.

BUDGET/P calculates the annual estimated financial statements and ratios for the entire bank. BUDGET/C, the model specification of which being almost similar to BUDGET/P, deals with each bank branch on a monthly basis. The budget variance analysis technique is also incorporated in the model.

A high degree of mathematical abstraction or sophistication is not involved in each SMART applications. The logic of the software tools for management should be as simple as possible for the managers to clearly understand. Otherwise, they are put off from utilizing the software tools by themselves. Therefore, complicated statistical techniques or sophisticated optimizing algorithms are not introduced in SMART.

3.2 BUDGET/P and BUDGET/C

As mentioned above, both of these two models are financial simulators, which can be used to test the alternatives (a set of decisions) in the decision making process. Optimization techniques, which are outstandingly successful in resource allocation problems, are not employed in the two models since they are often useless where objectives (objective functions) are complex as in real-life management.

The computational procedure, which is similar in the two models except that the calculation is performed monthly for each branch through BUDGET/C, is illustrated in Figure 4.

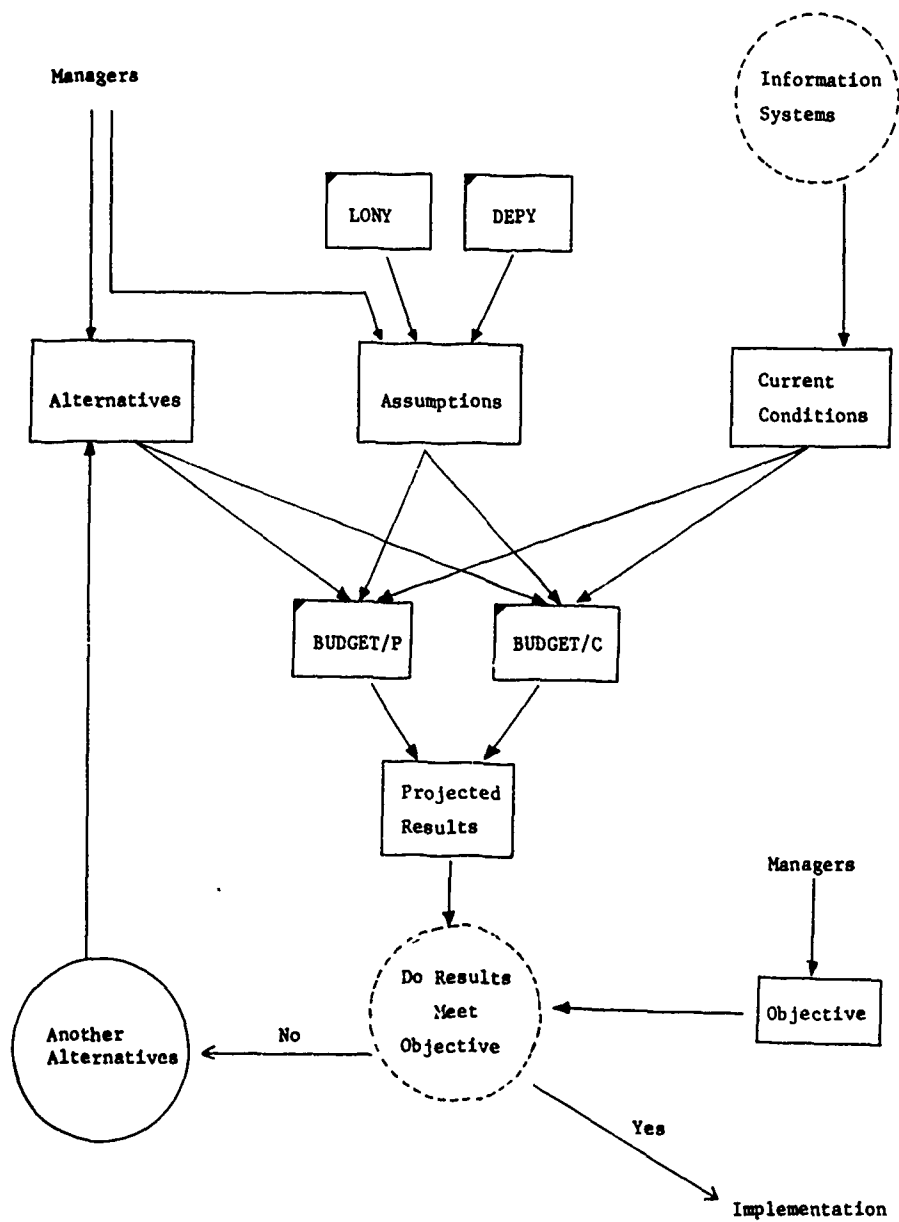


Figure 3 Decision Making Process and SMART

The primary output from the models consists of:

Projected Income Statements

Projected Balance Sheets

Projected Financial Ratios

Risk Analysis Results

Sensitivity Analysis Results

Iteration Results

BUDGET/P is designed to accomodate a three-year financial plan for the entire bank. On the other hand, BUDGET/C accomodates a twelve-month financial plan for each branch. The major functions of the models are as follows:

FORECASTING

The input forms are designed to permit the user to enter the key financial items needed to produce the financial forecasts. The forms can be considered as an effective communication media among the divisions related to the business planning.

SIMULATIONS

A major benefit derived from the model is the ability to perform "what if" analysis by changing the input data numerically representing the alternatives and assumptions. Three simulation analysis techniques, i.e., Risk Analysis, Sensitivity Analysis and Iteration are incorporated as integral parts of the models.

The Monte Carlo simulation is performed through the Risk Analysis module in the models. A triangular distribution function for the input items specified by the user is assumed and the calculated mean and standard deviations for the specified output items are reported.

One of the popular simulation techniques, Iteration, searches for the conditions which meet the management goals. The user specifies the output items and values related to the selected items which can be considered as a representation of the management goals. Then, the models

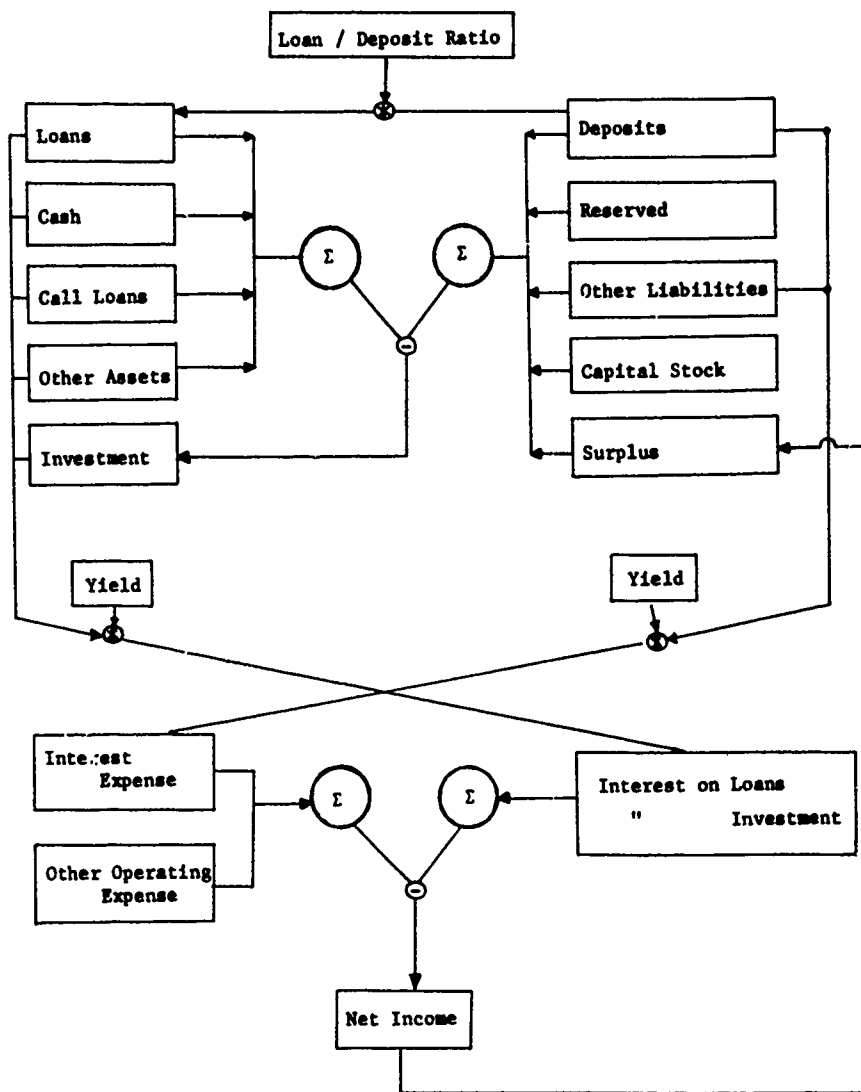


Figure 4 Flowchart of BUDGET/C

search for the conditions, i.e., values of input items which accomplish the specified objectives.

It is possible to perform parametric analysis through the Sensitivity Analysis technique, where the relative relationship between inputs and outputs are calculated and exhibited. Once the basic forecast has been made, the user can vary the planning items and test their effect on the other items. For example, one can quickly determine the effect of a Loan/Deposit ratio increase, on the income, yield on total assets, etc.

CONTROL

Normally, financial institutions have developed information systems designed to monitor the actual financial results and to compare them with the previous plans.

BUDGET/C can improve the above budgetary control by providing useful comparisons of projected financial results versus the actual ones. Based upon the Profit/Cost/Volume relationship in the model, it is possible to pinpoint the primary items where unfavorable variances occurred.

Furthermore, alternatives which can reduce the gaps between the planned and actual financial conditions are obtained through the "CONTROL" function of BUDGET/C.

3.3 DEPY AND LONY

The monthly estimated yield on commercial loans and average interest rate on deposits are obtained through the Loan Yield Model (LONY) and Deposit Yield Model (DEPY), respectively. Statistical techniques, such as least square method or exponential smoothing, are often employed in the estimation.

DEPY and LONY calculate the estimated yield not directly based upon these mathematical techniques but upon the current and forecasted rate-term mix of deposits and loans.

Figure 5 represents a computation of the yield on demand deposits.

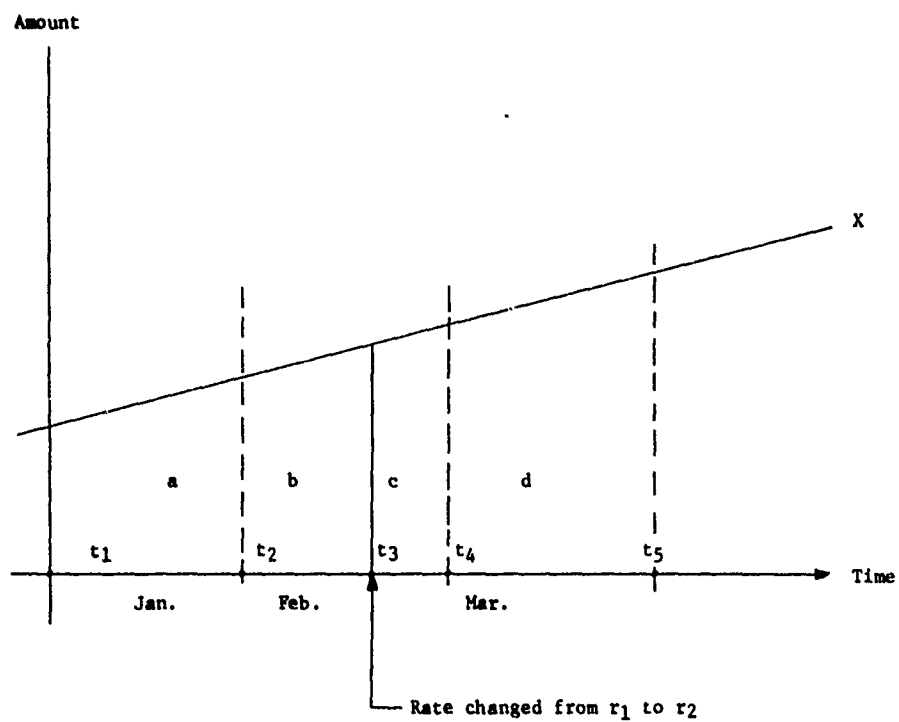


Figure 5 Yield on Demand Deposits

Curve x in Fig. 5 shows the estimated amount of demand deposit which is input to DEPY. The interest rate for the demand deposit is assumed to change from $Y_1\%$ to $Y_2\%$ at t_3 in February.

Area a is calculated as :

$$a = \int_{t_1}^{t_2} x(t) dt$$

Yield in January and March is r_1 and r_2 , respectively. However, the yield in February is computed as:

$$\frac{b*r_1 + c*r_2}{b + c}$$

Figure 6 represents the case for savings deposit where the interest rate is assumed to change four times in 1978. Curves x_1 , x_2 , x_3 , and x_4 which are estimated through DEPY denote the amount of each rate of savings deposit account.

The estimated yield in January '79, for example, is computed as :

$$\frac{5.25*a + 5.75*b + 6.00*c + 6.25*d + 7.25*e}{a + b + c + d + e}$$

Figure 7 represents the method of estimation for the yield on loans through LONY. The amount of loans in each month can be classified into two; one is the amount which is loaned in the current month and the other the amount which was previously loaned but still not withdrawn.

Characters a, b and c concerning the amount of loans in June (See Fig. 7) represent the balances which were loaned in March, April and May, respectively. Character d denotes the amount which was loaned in June.

Characters r_1 , r_2 , r_3 and r_4 represent the average interest rates for the loans in March, April, May and June, respectively.

The amounts for a, b and c are calculated based on the curves x_1 , x_2 and x_3 , respectively, which are estimated by LONY. The amount for d is input.

Yield on loans in June is calculated as:

$$\frac{r_1 \hat{a} + r_2 \hat{b} + r_3 \hat{c} + r_4 \hat{d}}{\hat{a} + \hat{b} + \hat{c} + \hat{d}}$$

where the marked (^) variables are estimated through the model.

The purpose of LONY and DEPY seems to be similar. However, the yield on loans and deposits are estimated in different manners. Contrary to the interest rates on deposits which are under the regulation of the government, interest rates on loans are market-determined and it is possible for each bank to change its rate-ratio based on its loan-policy.

In this sense, DEPY can be considered as a forecasting model and LONY as a policy-simulation model.

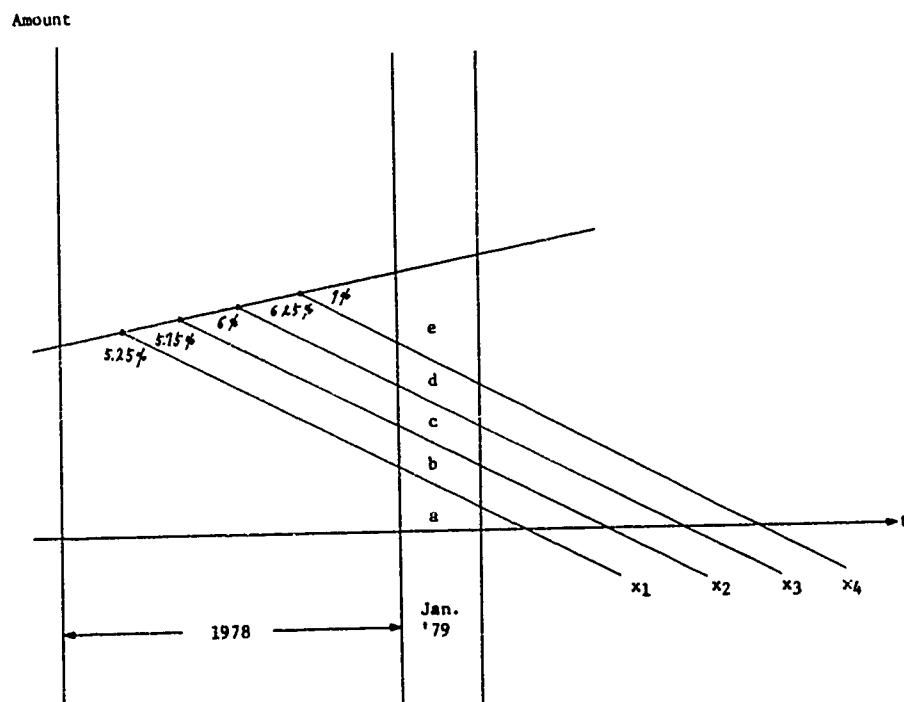


Figure 6 Yield on Savings Deposits

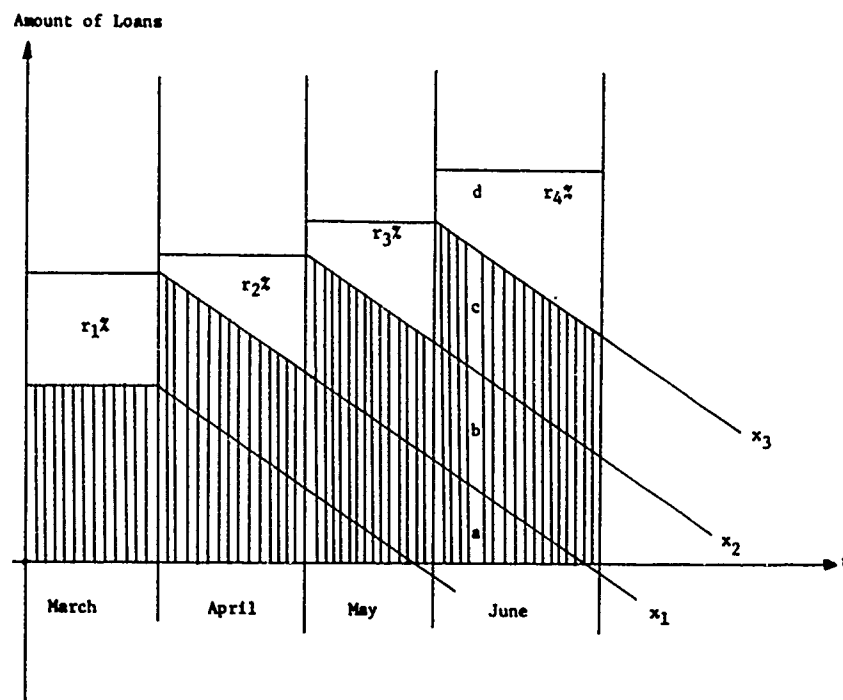


Figure 7 Yield on Loans

Conclusions

Our SMART application programs involved a long and continuous effort of the members of the project team in seeking some new techniques and developing actual software tools for solving practical business problems. In this project, we did not attempt to develop large scale or sophisticated applications but simple and easy to use tools for managers. With SMART requiring distinct definitions regarding the current business conditions, assumptions, and decision sets, important managerial informations can be obtained in a manner that is most useful for corporate planning.

We do not consider that the current version of SMART is perfect. However, basing on the experiences obtained from the actual installations, the design method employed in the development of SMART can be considered appropriate. At present, we devote to linking SMART with the so-called MIS in the financial institutions.

MINQUE Applied TO Regression Analysis

Moon Yul Huh

Statistician,
Software Development Center
Korea Institute of Science and Technology

Abstract

The purpose of this study is to investigate methodologies for estimating heterogeneous variances in linear models which plays a vital role to analyse a model when the experimentation or the observations do not come from the identical environment. This situation often appears in time phased data or when the experimentation is run in different circumstances. C.R. Rao's proposal (1970), named MINQUE, is a unified approach for this problem and this proposal is seen to be quite gratifying in this work. Also the often used method, weight least squares, is considered.

1. Introduction

Consider the linear model

$$Y = X \underline{b} + \underline{e} \quad \dots\dots\dots (1)$$

where \underline{y} is n -vector of observations with unexplainable portion \underline{e} having heteroscedastic variances. \underline{X} denotes $n \times p$ non-stochastic design matrix and \underline{b} is p -vector of unknown parameters.

In this work, we assume $\underline{e} \sim N(\underline{0}, D)$, where

$$D = \begin{pmatrix} \sigma_1^2 I_{n_1} & & \phi \\ & \sigma_2^2 I_{n_2} & \\ \phi & & \ddots \\ & & & \sigma_k^2 I_{n_k} \end{pmatrix}$$

$n_i \geq 1$ is the number of observations from the i -th group. There are considerable literatures (Ref. 6) on the estimation of \underline{b} in this case. When the parameters σ_i^2 ,

$i=1,2,\dots,k$ are known, the best linear unbiased estimator (BLUE) of \underline{b} is weighted least square estimator (WLSE) as given by $\hat{\underline{b}} = (X'D^{-1}X)^{-1}X'D\underline{y}$. However seldom is D known, and good estimate of D naturally leads to good estimate of \underline{b} . Sometimes estimating D itself can be valuable. This is the case when the stratified sampling is of concern rather than simple random sampling.

Recently Rao (1970) has suggested a method known as MINQUE to estimate components of variance in a general linear model.

2. The MINQUE Theorem

To apply the methodology to our problem, rewrite the model (1) as

$$Y = X \underline{b} + \underline{q}_1 + \underline{q}_2 + \dots + \underline{q}_k$$

where $\underline{q}_i \sim N(\underline{0}, \sigma_i^2 V_i)$

and

$$V_i = \begin{pmatrix} \phi & \vdots & \phi & \vdots & \phi \\ \phi & \sigma_i^2 I_{n_i} & \phi \\ \phi & \vdots & \phi & \vdots & \phi \end{pmatrix}$$

Denoting

$$R_* = D^{-1} - D^{-1}X(X'D^{-1}X)^{-1}X'D^{-1}, \quad \text{Rao (1970)}$$

showed that MINQUE of $\underline{\delta}' = (\sigma_1^2, \dots, \sigma_k^2)$ is

$$S_* \hat{\underline{\delta}} = \underline{u}_*$$

where

$$S_* = \{ \text{tr}(R_* V_i R_* V_j) \}$$

$$\underline{u}_* = \{ Y'R_* V_i R_* Y \} \quad i, j=1, 2, \dots, k.$$

and trace (A) denotes the trace of a rectangular matrix A. However D is usually unknown and a-priori values instead of D is suggested. Then the MINQUE estimate, $\hat{\underline{\delta}}$, of $\underline{\delta}$ is obtained from solving the following system of linear equations.

$$S \hat{\underline{d}} = \underline{u},$$

$$\text{where } S = \left\{ \text{tr}(RV_i RV_j) \right\}$$

$$\underline{u} = \left\{ \underline{y}' RV_i R \underline{y} \right\}, \quad i, j = 1, 2, \dots, k.$$

$$R = W - WX(X'WX)X'W$$

$$\text{and } W = D^{-1} \text{ given a-priori values of } \sigma_i^2.$$

3. PROPERTIES OF MINQUE.

The following important properties of MINQUE in general linear model are given without proof. Interested reader is suggested to read Rao(5) and Huh(2)

- i) MINQUE is unbiased regardless of the a-priori values.
- ii) MINQUE is invariant under the transformation of unknown parameters
- iii) MINQUE is minimum variance quadratic unbiased estimator under normality.

Further, noting that R appears twice in either side of the estimating equation (2-2), MINQUE with a-priori values proportional to the true parameter values is the same as MINQUE with correct a-priori values.

The problem associated with MINQUE is

- i) MINQUE may yield negative estimates.
- ii) Computationally, MINQUE is not easy.
- iii) MINQUE requires a-priori values.

Several papers (1,6,7) were devoted for the problem of i) and some modifications have been suggested by them. For the 2nd problem, Lon Liu Ku and Senturia (4) have suggested a simpler form. Also Huh has alleviated this problem greatly in the estimation of variance components of two-way random model without interaction. However, the biggest problem is seen to be determining a-priori values. When we don't have any a-priori knowledge of the underlying circumstances, the easiest way to implement a-priori knowledge would be to assign equal weights to all parameter values as if the observations were from a homogeneous system. Hence it is important to investigate how close the estimator obtained in this way is close to

the optimum estimator, optimum in the sense of minimum variance estimator among the quadratic unbiased estimators. The most common approach to evaluate an estimator is via the mean square error. Hence we define the efficiency of an estimator as in the following.

$$\begin{aligned} & \text{efficiency of an estimator} \\ &= \frac{\text{MSE of the estimator}}{\text{MSE of MINQUE with correct a-priori values}} \end{aligned}$$

4. Variances of the Estimator

The variance of MINQUE is obtained as

$$\text{var}(\hat{\underline{\beta}}) = \text{var}(\underline{S}^{-1}\underline{u}) = \underline{S}^{-1} \text{var}(\underline{u}) \underline{S}^{-1}$$

$$\begin{aligned} \text{But } \text{var}(\underline{u}) &= \left\{ \text{cov}(u_i, u_j) \right\} \\ &= \left\{ \text{cov}(\underline{y}' \underline{R} \underline{V}_i \underline{R} \underline{y}, \underline{y}' \underline{R} \underline{V}_j \underline{R} \underline{y}) \right\} \\ &= \left\{ 2 \text{tr}(\underline{R} \underline{V}_i \underline{R} \underline{D} \underline{R} \underline{V}_j \underline{R} \underline{D}) \right\} \end{aligned}$$

This is very difficult to visualize and analytical investigation of the above quantity is seen to be far from tractability. We restrict our attention only for the case of $p = 1$. After some manipulations, the variance-covariance matrix of \underline{u} is obtained as

$$\begin{aligned} \text{Cov}(u_i, u_j) &= \\ &\left\{ \begin{array}{l} 2/(r_i^8) \left[(n_i-1)\sigma_i^4 + (\sigma_i^2 + m_{ii}x_i'x_i)^2 \right] \\ \text{when } i=j, \text{ and when } i \neq j \text{ is} \\ 2m_{ij}^2 x_i'x_i x_j'x_j / (r_i^4 r_j^4) \end{array} \right. \end{aligned}$$

and the elements of $k \times k$ matrix S is given as

$$s_{ij} = \begin{cases} (n_i - 1) / r_i^4 + (1 - \underline{x}_i' \underline{x}_i / r_i^2)^2 / (l^2 r_i^4) & , i=j \\ \underline{x}_i' \underline{x}_i \underline{x}_j' \underline{x}_j / (l^2 r_i^4 r_j^4) & , i \neq j \end{cases}$$

where

$\underline{x}' = (\underline{x}_1', \underline{x}_2', \dots, \underline{x}_k')$,
 \underline{x}_i is n_i -vector of constants, r_i^2 is the a-priori values of

$$l = \sum_{i=1}^k \underline{x}_i' \underline{x}_i / r_i^2 ,$$

$$m_{ii} = \underline{x}' W D W \underline{x} / l^2 - \sigma_i^2 / (r_i^2 l) - \sigma_j^2 / (r_j^2 l)$$

Now the variance of often used estimator, $s_i^2 = \underline{y}_i' (I - \underline{x}_i (\underline{x}_i' \underline{x}_i)^{-1} \underline{x}_i') \underline{y}_i$, is $2 \sigma_i^2 / (n_i - 1)$

Note that s_i^2 is impossible to obtain when $n_i = 1$ whereas MINQUE is possible even with $n_i = 1$.

5. Numerical Investigation

Factors governing the variances of MINQUE seem to be n_i , \underline{x}_i , r_i^2 , σ_i^2 for $i=1, 2, \dots, k$ and k . Since it is far from practicality to investigate all possible ranges of the factors, $n = \sum_{i=1}^k n_i$ was chosen to be 10, 20, 40. Also $k = 3, 5, 7$ and range of were chosen only in the range of 1. through 10. Two sets of a-priori values of r_i^2 were chosen as in the following:

i) All a-priori values are chosen proportional to the true value, i.e., $r_i^2 = c \sigma_i^2$ for all i and for any positive constant c .

ii) All a-priori values are chosen as unit.

These choices are because the first one yields the optimum estimators and second one is suggested by C.R. Rao when the analyst does not have any a-priori knowledge.

Most of the numerical results can be seen from the two tables in the Appendix.

Conclusion

This work has shown the following results:

i) MINQUE is quite robust for almost the cases investigated.

ii) For almost the cases investigated the s_1^2 was seen be about 50% worse than the optimum estimator.

The first result is quite gratifying and this is nice property of MINQUE because it gives estimates very close to the optimum one regardless of the many factors mentioned earlier in section 5. Also in many practical situations, it is usually the case that relative weights of the stratum variances are known. This will make the MINQUE solution optimum. Even when no a-priori knowledge is available, MINQUE yields quite satisfactory results. Hence MINQUE is highly recommended for any circumstances and this estimates can be applied to weighted least squares method to analyse the linear model considered in (1).

References

1. Horn, S.D., Horn, R.N., and Duncan, D.B., "Estimating Heteroscedastic variances in Linear Models", Journal of the American Statistical Association 70, (1975), 380-385.

2. Huh, M.Y. "Robustness of the MINQUE Procedure in Estimating Variance Components", Unpublished Dissertation, Department of Statistics, Southern Methodist University, (1978)
3. Jacquez, J.A. and et al., "Linear Regression with Non-constant, Unknown Error Variances : Sampling Experiments with Least Squares, Weighted Least Squares and Maximum Likelihood Estimators", Biometrics 24, (1968), 607-626.
4. Liu, L.M. and Senturia, J., "Computation of MINQUE Variance Component Estimates", Journal of the American Statistical Association 72, (1977), 867-868.
5. Rao, G.R., "Estimation of Heteroscedastic Variances in Linear Models", Journal of the American Statistical Association 65, (1970), 161-172.
6. Rao, J.N.K. and Subrahmaniam, k., "Combining Independent Estimators and Estimation in Linear Regression with Unequal Variances", Biometrics 27, (1971), 971-990.
7. Rao, P.S.S.R., "Theory of MINQUE Review", Discussion Paper No. 118, Indian Statistical Institute, New Delhi, (1975).

Appendix

In the following 2 tables 2 typical results are given. In table 1, K is 3 and table 2 shows the results when $K=7$. $\text{Var}(\hat{\sigma}_i^2)$ stands for the variance of the optimum estimator, while $\text{Var}(\hat{\sigma}_i^2)$ denotes the variance of the MINQUE of σ_i^2 when all the a-priori values for the parameters are set to 1. $\text{Var}(S_i^2)$ is the variance of the often used estimator of the variance of i -th stratum. The numbers in parenthesis is the values when the η_i 's are doubled.

T A B L E 1

η_i	2	4	4
σ_i^2	7.0	5.0	9.0
$\text{var}(\hat{\sigma}_i^2)/\sigma_i^2$	1.014 (0.514)	0.585 (0.286)	0.643 (0.250)
$\text{var}(\hat{\sigma}_i^2)/\text{var}(\hat{\sigma}_i^2)$	1.001 (1.000)	1.017 (1.000)	1.012 (1.000)
$\text{var}(\hat{\sigma}_i^2)/\text{var}(\hat{\sigma}_i^2)$	1.972 (1.296)	1.140 (1.000)	1.037 (1.143)

T A B L E 2

η_i	4	1	2	2	3	2	7
σ_i^2	1.0	8.0	6.0	9.0	3.0	2.0	2.0
$var(\hat{\sigma}_i^2)/\sigma_i^4$	0.506 (0.267)	2.010 (1.033)	1.011 (0.537)	1.013 (0.543)	0.709 (0.370)	1.152 (0.500)	0.329 (0.143)
$var(\hat{\sigma}_i^2)/var(\hat{\sigma}_i^2)$	1.002 (1.054)	1.001 (1.016)	1.002 (1.034)	1.002 (1.046)	1.012 (1.063)	1.028 (1.000)	1.023 (1.000)
$var(\hat{\sigma}_i^2)/var(\hat{\sigma}_i^2)$	1.318 (1.072)	Not available (1.937)	1.978 (1.241)	1.975 (1.228)	1.410 (1.082)	1.736 (1.333)	1.012 (1.077)

THE AMMUNITION STOCKPILE
RELIABILITY PROGRAM

ALAN S. THOMAS

ROBERT M. EISSNER

Reliability, Availability and Maintainability Division
US Army Materiel Systems Analysis Activity
Aberdeen Proving Ground, Maryland 21005

ABSTRACT. The Ammunition Stockpile Reliability Program (ASRP) is a life cycle commodity oriented world wide logistics support program conducted to ascertain the reliability, safety and performance characteristics of stockpiled and deployed ammunition systems. The ASRP provides the US Army with sound objective information for logistical decisions regarding the storage, maintenance, modification, retention, replacement, supply and use of ammunition. Included in the ASRP are conventional, toxic chemical, nuclear and missile ammunition items.

The ASRP testing program consists of both functional testing (ballistic testing at proving grounds and other field testing facilities) and laboratory testing (destructive or non-destructive testing for physical/chemical/electronic analysis). The program is cyclic in structure as well as being a sequential stratified random sampling program. AMSAA conducts a significant portion of the ASRP for conventional ammunition.

Based upon individual ammunition item design, storage, handling, and use requirements, and the quality level demonstrated when accepted; definitive sampling plans, testing procedures, classification of defects, lot quality standards and grading criteria are established for each item included in the program. Subsequent to the analysis and evaluation of the individual item cyclic test, a test report providing background information, test results, conclusions and recommendations concerning the item stockpile is published and distributed by AMSAA.

The objective of the Ammunition Stockpile Reliability Program (ASRP) is to assure that the ammunition the U. S. Army has in its stockpile meets all of its safety and reliability requirements. To accomplish this objective the functions as shown below must be accomplished.

- o Monitor stockpile quality
- o Detect unsatisfactory conditions/trends
- o Investigate malfunctions
- o Restrict/suspend unsatisfactory munitions
- o Identify items for renovation or disposal

In performing these functions it is the purpose of the ASRP to evaluate the safety, serviceability, reliability and performance of the ammunition items in stockpile or deployed in the hands of troops and assure that these items are in a state of readiness to perform reliably and effectively upon demand, with the shortest of notice, anywhere in the world. At the same time the ASRP can provide sound, technical information for decisions on phasing of replacements, renovation, maintenance and supply of ammunition in order to minimize the costs of storing unsafe, unreliable, unserviceable ammunition.

The classes of materiel covered by the Stockpile Reliability Program are as shown here.

- o Guided missiles and large rockets
- o Chemical ammunition
- o Chemical protective equipment
- o Artillery, armor and infantry ammunition
- o Propellant charges and bulk propellant
- o Small arms ammunition
- o Nuclear weapons
- o Explosive loaded components
- o Mines, grenades, simulators, signals and firing devices

Because of the basic differences in all these classes of materiel - differences in design, performance and use along with a mixture of low density, high cost items as opposed

to high density, low cost items - the structure and responsibilities of the Stockpile Reliability Program differ depending on the item. However, there is a basic structure common to all items and after first showing who has what responsibilities in the program, this basic structure will be discussed followed by a slightly more detailed account of the program for one of the classes of materiel.

Figure 1 shows the organizational responsibilities from the Department of the Army in the person of the Deputy Chief of Staff for Logistics (DCSLOG) and the Deputy Chief of Staff for Research, Development and Acquisition (DCSRDA) down through the U. S. Army Development and Readiness Command (DARCOM) to the major Army commands (MACOM).

Responsibilities

Organizations

- o Department of the Army
 - DCSLOG - Logistics
 - DCSRDA - Acquisition
- o DARCOM
 - DRCQA Program Control
 - ARRCOM
 - MIRCOM Conducts Program
 - PM NUC
 - ARRADCOM
 - MIRADCOM Provides support
 - DESCOM
 - TECOM
 - Other
 - AMSAA Performs Independent Assessments
- o MACOMS Field Surveillance Operations

Figure 1 - Organizational Responsibilities

Within DARCOM the Director of Quality Assurance (DRCQA) has overall program control while the readiness commands - Armament Readiness Command (ARRCOM) and Missile Readiness

Command (MIRCOM) - conduct the programs for conventional ammunition items and missiles. The Project Manager for Nuclear Weapons conducts the nuclear program.

The research and development commands - Armament (ARRADCOM) and Missile (MIRADCOM) - provide technical and engineering support; the Depot Support Command (DESCOM), besides storing the ammunition, performs the inspections and does testing of those items suitable for depot testing; the Test and Evaluation Command (TECOM) performs the testing of those items requiring proving ground tests; and the Army Materiel Systems Analysis Activity (AMSAA) provides a technical review and performs an independent assessment of the overall program for DARCOM.

The program execution consists of the following elements:

- o Using unit surveillance
- o Army depot surveillance operations
- o Laboratory testing
- o Function testing
- o Malfunction activities

In executing this program it might be said the surveillance program determines the non-functional condition of the ammunition whereas the stockpile test program determines the function condition of the ammunition.

The surveillance portion of the ASRP is performed by Quality Assurance Specialists (Ammunition Surveillance) who provide logistical support to units in the field in the form of annual basic load inspections, periodic inspections of ammunition supply points to check for deterioration and damage, malfunction investigations and explosive safety in support of ammunition operations. At depots the QA specialists perform inspections and prepare reports as shown in Figure 2.

Receipt Inspection

- o Visual - Mechanical (gage) for damage in transit and gross manufacturing defects

Periodic Inspections

- o Visual - mechanical - electrical
- o By lot or serial number (missiles) for deterioration
- o Special tests

Storage Monitoring

- o Checks for leakers
- o Controlled humidity containers

Special Inspection

- o Command directed

Issue Inspection

- o Confirm serviceability

Reports

- o DA Form 2415 - Ammunition condition reports
- o DA Form 984 - Munitions surveillance reports
- o Ammunition inspection and lot report

Figure 2 - Surveillance Program at Depots

Laboratory tests are conducted for many reasons and are used in support of visual inspections and functional tests or in some instances as the only method of evaluating the quality of items in the stockpile. They include chemical analysis in a laboratory for determination that chemical compositions of explosives, propellants or chemical fillers have not changed or deteriorated as well as ultrasonics, electrical checks, etc.

Function testing may be conducted as part of a depot function test, proving ground tests, or training and annual service practice (ASP) firings. These tests may be conducted as part of a planned, stratified program for a particular item or as one step in a malfunction investigation.

Depot tests are normally limited to small items such as anti-personnel mines, pyrotechnics, and grenades and are controlled by ARRCOM under the Centrally Controlled Function Test (CCFT) Program. ARRCOM also controls and funds for the proving ground tests as recommended and performed by AMSAA or the responsible engineering agency. Proving ground tests normally involve large ammunition items such as artillery ammunition, anti-tank mines, mortars and small rockets which require special test facilities and equipments. Field practice firings are usually conducted on missile systems. These tests may be instrumented missiles under controlled conditions or unit firings of stockpiled missiles.

Having given the basic structure of the program a little more detailed discussion of the philosophy and mechanics of one of the portions of the program - the large caliber ammunition proving ground test program conducted for ARRCOM by AMSAA - is now in order.

The basic provisions of this program are as shown in Figure 3.

Criteria

- o Original product baseline
- o Quality required for effective service use
- o Safety

Inspection and Test Attributes

- o Storage inspection & laboratory testing
- o Extreme ballistic conditions
- o Quality provisions conditions
- o Expected use in the field
- o Past experiences - failure mode

Sampling

- o Experimental design
- o Sequential & stratified random sampling program
 - storage location, climatological areas
 - manufacturers, age, lot size, history

Evaluation

- o Estimate stockpile reliability
- o Individual lot grade
- o Three courses of action - retain - use now - renovate
- o Grading criteria - Minimize sum of probability of misgrading

Figure 3 - Stockpile Reliability Provisions

No discussion of these provisions will be given right now as each of the provisions will be discussed during a brief description of the mechanics of the program.

As to the mechanics, the starting point for the program is the establishment of a five-year plan. The nature of this five-year plan is as shown in Figure 4.

Five-Year Plan

1. Lists items, number of lots and total quantities to be tested by year for a five-year period.
2. Program is a sequential stratified random sampling type program - test as much as often as need be (starting with as small a number of lots as possible) to evaluate the items.
3. Coordinated with various ARRCOM directorates and subordinate organizations. Also coordinated with TECOM pertaining to the use of their facilities.
4. Updated annually
5. Items selected for test depend on
 - a. Inventory - availability and size of stockpile
 - b. Previous surveillance test results - the sequential nature of the program
 - c. Field results - malfunctions

Figure 4 - Five-Year Plan

The key points here are the sequential nature of the program and the coordination with all the elements of ARRCOM.

Once any given years program is established, the next stage is the sampling of the stockpile. Stocks are sampled from all over the world and shipped back to a U. S. proving ground for test. The number of lots sampled from the worldwide stockpile inventory depends on the stratification of the stockpile, the results from previous tests and the experimental design to be used. The stockpile is stratified according to:

- o Storage location
- o Climatological areas
- o Manufacturers
- o Age

- o Lot sizes
- o Models

Once the samples are received at a proving ground, the lots are tested in accordance with a given test directive. The test directive gives the test phases to be conducted, the sample sizes per lot for each test phase, the observations and measurements to be taken and the methods of test and equipment to be used. The tests are usually conducted at the extremes of ballistic conditions, the acceptance test conditions for comparison purposes, and the expected conditions most used in the field if different from those previously mentioned. It's important to mention here that the tests are conducted using some kind of experimental design to increase the precision of the inferences made from the test, facilitate reduction of the data and reduce the costs in conducting the tests.

The individual lots tested are then assigned one of four grades in accordance with objective criteria which specify lot quality standards, classification of defects and grading criteria. The lot quality standards are based on determination of the quality that the majority of reliable producers were able to produce and the quality required for effective service use.

Stockpile evaluations are then performed, functional codes recommended so that more positive action with regard to the individual lots can be taken, and recommendations pertaining to the entire stockpile are given - whether the stockpile is satisfactory, certain strata are bad, more testing is required, restrictions should be employed, etc.

The results of these evaluations are published in two types of reports. One is a compilation of individual item test reports published periodically; the second is a once a year executive summary providing an overview of the status of each of the conventional ammunition items in the stockpile.

On receiving the individual item test reports ARRCOM takes the recommendations and performs engineering and supply evaluations to arrive at the final disposition of the stockpile. These engineering and supply evaluations are performed to check into the disposition instructions for the munitions, to see if any change in the technical data package is required, to see if any engineering investigations or development programs are required, to

ascertain the relationship of the results to similar items, to establish the logistical impact of the recommendations and to make priority assignments to the recommendations.

To give an idea what the stockpile reliability program for a given ammunition item might consist of, the visual inspection program and the proving ground function test programs for the 105mm HE, M444 round is outlined in Figure 5.

Visual Inspection of Munitions

- o Inspection interval - 4 years
- o Receipt inspection - for damage in transit and gross mfg defects
- o Periodic inspections - deterioration defects

Proving Ground Function Test

- o Post production interval - 3 years
- o Visual inspection
- o Mechanical (gage)
- o Lab analysis of propellant
- o Firing, performance and safety phases include:

Muzzle Velocity	Submissile Performance	Height of Burst
Pressure	Precision	Projectile Reli-
Fuze Performance	Post-Mortem Analysis	ability
Time of Flight	Range to burst	Extremes of
		Ballistic
		Condition

Figure 5 - 105mm, HE, M444 Stockpile Reliability Program

Having thus conducted evaluations as described on all the ammunition items, it is readily apparent there are many important feedbacks that can be obtained from the Ammunition Stockpile Reliability Program. The feedbacks that can be obtained from this program are shown in Figure 6.

- R&D - Life cycle characteristics
- Engineering - Design changes
- Procurement - Tech Data Package Reprocurement
- Stockpile Management - Priority of Issue
 - Retention, Disposition
- Supply - Distribution vs. Demands
- Maintenance Program - Retrofit
- Dissemination of information to field via changes to TB's, TM's, letters, etc.
- Systems Analysis - Data base
 - New requirements
- Firing table corrections

Figure 6 - Feedback from ASRP

Thus in addition to having a program to assure the readiness and performance capabilities of the stockpile, the ASRP feeds back valuable information to R&D and procurement to help assure better ammunition coming into the stockpile.

ARTILLERY FORCE SIMULATION MODEL (AFSM)

ALAN S. THOMAS

RICHARD S. SANDMEYER

US Army Materiel Systems Analysis Activity
Aberdeen Proving Ground, MD U.S.A. 21005

ABSTRACT. The performance in combat of a division slice of Blue Artillery is a function of its weapon-ammo basic load and resupply, fire direction center (FDC) capability, movement policy, firing policy, weapon reliability, and weapon repair capability as well as red anti-artillery capabilities such as counterbattery acquisition systems, counterbattery fire capability and doctrine, and electronic warfare capability. These factors are inputs to AFSM and may be varied from run-to-run to determine the effect of each factor on the performance of BLUE Artillery.

AFSM is a non-dynamic weapons effectiveness model which calculates the damage that a BLUE Artillery force could do to a given RED threat force when the factors listed above are taken into account. It is non-dynamic because the list of RED units acquired by BLUE as potential artillery targets is predetermined by an externally played wargame (such as DIVLEV or DIVWAG) and is not varied within AFSM.

The output of AFSM includes such measures of effectiveness as RED personnel, tanks, APC's, trucks, artillery tubes, radar and missile launchers killed by BLUE artillery as well as measures of BLUE artillery's efforts such as ammo fired (broken down by round types and ranges), time used to process and fire missions by the FDC's and batteries, BLUE artillery tubes lost due to reliability, and BLUE tubes lost due to attrition by RED counterbattery fire. The artillery tube losses from reliability and attrition are combined with the maintenance and repair capability of the force to derive an estimate of artillery battlefield availability as a function of battle time.

The results from several AFSM runs are compared to show the changes caused by varying certain inputs.

OUTLINE

1. INTRODUCTION

2. MODEL DESCRIPTION

2.1 Model Events

2.2 Target Tape

- 2.2.1 Fire Mission
- 2.2.2 Meteorological (MET) Message
- 2.2.3 Artillery Target Intelligence (ATI) Report
- 2.2.4 Surveys
- 2.2.5 Fire Plans

2.3 AFSM Inputs (Other Than Target Tape)

- 2.3.1 Weapon System Descriptors
- 2.3.2 Artillery Round Attributes
- 2.3.3 Posture Sequences
- 2.3.4 RED Electronic Warfare Schedule
- 2.3.5 BLUE FDC Processing Times
- 2.3.6 BLUE Battery Movement Schedule
- 2.3.7 RED Counterbattery-Counter mortar Radar Data
- 2.3.8 BLUE Required Damage Levels
- 2.3.9 RED Counterbattery Volume of Fire
- 2.3.10 BLUE Artillery Tactical Assignments
- 2.3.11 RED Artillery Force Organization

2.4 BLUE FDC Events

- 2.4.1 BLUE Fire Mission
 - 2.4.1.1 At the FDC
 - 2.4.1.2 Within Each Battalion
 - 2.4.1.3 CLGP Fire Mission
 - 2.4.1.4 GSRS
 - 2.4.1.5 Missiles
- 2.4.2 MET Messages
- 2.4.3 ATI Reports
- 2.4.4 Survey
- 2.4.5 Fire Plans
- 2.4.6 Other BLUE FDC Events

2.5 BLUE Battery Events

- 2.5.1 Firing a Fire Mission
 - 2.5.1.1 HE and ICM
 - 2.5.1.2 CLGP
 - 2.5.1.3 Updating Damage to RED Units

OUTLINE (CONTINUED)

- 2.5.1.4 RED Counterbattery Target Acquisition
 - Radar
 - Sound
 - Moving Target Intelligence
- 2.5.1.5 Scheduling RED Counterbattery Fire
- 2.5.2 Firing a Fire Plan Target Mission
- 2.6 RED Counterbattery Fire Event
- 2.7 RED Electronic Warfare
- 2.8 Defeat
- 2.9 Suppression
- 2.10 Mini-Moves
- 2.11 Tube Losses and Repair
- 2.12 AFSM Output
 - 2.12.1 BLUE Artillery Force Performance
 - 2.12.2 BLUE Round Expenditures
 - 2.12.3 BLUE Artillery Busy Times
 - 2.12.4 Range Tables
 - 2.12.5 Unaccomplished Mission Summary
 - 2.12.6 Reliability and Attrition Results
 - 2.12.7 BLUE System Performance Table
 - 2.12.8 BLUE GSRS Performance Table
 - 2.12.9 RED Fire Unit Status Table
 - 2.12.10 RED Radar Unit Status
- 3. SAMPLE RESULTS
 - 3.1 Mix Descriptions
 - 3.2 Results
 - 3.2.1 BLUE Force Effectiveness
 - 3.2.2 BLUE Force Measures of Effort

1. INTRODUCTION

The Artillery Force Simulation Model (AFSM) is a division level, nondynamic, artillery weapons system effectiveness model.

The model evaluates the performance of a division slice of BLUE artillery considering its weapon-ammo mix, ammo basic load and resupply rate, fire direction center (FDC) capabilities, movement doctrine, firing doctrine, weapon reliability, and weapon repair capability as well as RED's anti-artillery capability as represented by RED counterbattery target acquisition devices, RED counterbattery fire, and RED electronic warfare (EW).

AFSM considers the factors mentioned above as it simulates the actions of a BLUE artillery force attempting to satisfy demands placed on it. These demands mainly take the form of artillery fire missions against a list of RED units acquired as potential artillery targets. This list is obtained from the playing of an external wargame (such as DIVLEV or DIVWAG) and consists of a time ordered list of RED units acquired by BLUE as potential artillery targets. AFSM is termed non-dynamic because with one exception explained below, the RED force maneuvering, as represented by the acquisitions on the target list, is independent of the performance of the BLUE artillery force.

2. MODEL DESCRIPTION

2.1. Model Events

AFSM is mainly an event sequenced model with three major event types: BLUE FDC events, BLUE battery events, and RED counterbattery fire events. In addition, there are two events that occur at regular intervals: the quarter-hourly tube status check and the hourly game summary print-out.

2.2. Target Tape

An external wargame such as DIVWAG or DIVLEV is played with a given force-on-force scenario and during the game, a record of target acquisitions made by the BLUE force is kept. This list of target acquisitions is analyzed and the functions of a Tactical Operations Center (TOC) are simulated to produce a list of the requests for BLUE artillery fire that would result from such a list of target acquisitions.

To these requests for BLUE artillery fire (hereafter called fire missions) are added four other BLUE FDC event types: meteorological (MET) messages, artillery target intelligence (ATI) reports, surveys, and fire plans. The resulting list of BLUE FDC events is called the target tape.

Each event on the target tape is associated with a certain BLUE FDC and the events are ordered in time sequence by their respective initiation times at their FDC's.

2.2.1 Fire Mission

A fire mission is a request directed to a particular BLUE FDC for artillery fire against a target. It includes:

(a) estimated target descriptors such as target radius, target unit type, target military worth (used to establish target priority), target cover type (i.e., environment), target departure time from current location, target posture.

(b) actual target descriptors such as target radius, target unit type, target cover type, target departure time from current location; target posture.

(c) a target location error (TLE) associated with the acquisition.

(d) coordinates of target center

(e) fractions of the target unit's original personnel, tanks, APC's, etc. that have already been killed by BLUE non-artillery weapons (maneuver units, tactical air) up to the current acquisition.

(f) for those forward observer (FO) acquired targets that are thought suitable for Copperhead (155mm cannon launched guided projectile) fire, a CLGP window (time interval during which the FO could designate the target with a laser) is included.

The estimated target descriptors are used to decide what, when, and how to fire at the target. The actual target descriptors are used to determine the effects achieved against the target when rounds are actually fired.

2.2.2 Meteorological (MET) Messages

The target tape includes meteorological (MET) messages among its events. A MET message in AFSM has no effect

other than to make the FDC busy for a brief time. This is supposed to simulate the load on the FDC that results from entering the new MET data into the FDC computer.

2.2.3 Artillery Target Intelligence (ATI) Reports

Certain target acquisitions are of duration too short for requesting artillery fire. These acquisitions become ATI reports and in AFSM they are played only by charging the FDC some processing time to record the ATI report in its computer memory.

2.2.4 Surveys

During the course of the battle, artillery batteries make moves from one site to another. Before moving into a new site, it is desirable (in order to keep delivery errors low) to have an accurate survey of the new site. The survey missions in AFSM merely charge the FDC some processing time to take into account the time required to do the survey calculations.

2.2.5 Fire Plans

A fire plan as played in AFSM is a list of battlefield areas onto each of which it is desired to deliver a certain amount of artillery fire at a given time (or within a specified time interval.)

In AFSM a fire plan consists of a message sent to an FDC about forty-five minutes prior to the desired fire time informing it of the areas to be fired on, the times of the fires, and the volume of fire in terms of volleys of 155mm HE rounds.

2.3 AFSM Inputs (Other than Target Tape)

The target tape, once generated by the external wargame and modified to include MET, ATI, survey, and fire plan missions, is fixed for all runs made which consider the given RED threat scenario.

The other inputs, which are varied to study the effects of different weapon-ammo mixes, force compositions, etc., are described in this section.

2.3.1 Weapons System Descriptors (For Both BLUE and RED Artillery)

- (a) weapon system ID code
- (b) range
- (c) rates-of-fire
- (d) basic load and resupply rate for ammunition
- (e) number of tubes (or launchers) per battery
- (f) mean rounds between various types of firepower failures
- (g) mean kilometers travelled between various types of mobility failures
- (h) repair times for the various types of failures as well as for various types of attrition damage
- (i) maximum number of volleys to be fired by any one battery on a given mission.

2.3.2 Artillery Round Attributes (For Both BLUE and RED)

- (a) round ID code
- (b) round crated weight
- (c) round cost
- (d) round maximum range
- (e) round in flight reliability
- (f) basic load and resupply rate of the specific round
- (g) lethal areas vs various target elements (e.g., standing personnel, prone personnel, crouching personnel, tanks, APC's, trucks, artillery tubes, etc.) at various ranges in various cover types
- (h) for ICM rounds only: submunition pattern radius, number of submunitions, and reliability of submunitions

(i) for Copperhead (CLGP) lethal areas are replaced by a table giving expected tank (or APC) kills as function of number of rounds fired

(j) round delivery errors (both MPI and precision).

2.3.3 Personnel Posture Sequences for Various Target Types

(a) fraction of target unit personnel in each of standing, prone, and crouching-in-foxhole postures prior to warning

(b) same after warning (i.e., after becoming aware of incoming fire)

2.3.4 Times during game that RED commences and discontinues Electronic Warfare.

2.3.5 Processing times for various BLUE FDC tasks (varies with FDC computer type).

2.3.6 Battery Movement Information

(a) arrival and departure times for each of battery's sites during game (for BLUE and RED)

(b) map coordinates for each of battery's sites during game (for BLUE and RED)

(c) number of incoming fires a BLUE battery will endure at a site before making a mini-move

(d) number of volleys a BLUE battery will fire from a site before making a mini-move

2.3.7 RED Counterbattery-Countertermortar Radar Data

(a) radar unit movement schedules and site coordinates

(b) radar's target acquisition probabilities as a function of radar-to-battery range, type of round being tracked, and radar's simultaneous tracking ability

(c) target location error (TLE) factor associated with radar type

(d) radar reliability data (mean time between failures and estimated repair time)

2.3.8 Damage levels desired against various target types (used in deciding how many rounds BLUE will fire against a given target)

2.3.9 RED counterbattery volume of fire tables used to determine number of rounds RED will fire on a counterbattery mission.

2.3.10 Tactical Assignments for BLUE Artillery Units

A BLUE artillery battalion may be assigned one of six roles:

(a) Direct Support. This means it is supporting a maneuver unit (usually a brigade) and gives first priority to requests for fire from that brigade.

(b) Reinforcing. In this role, the battalion reinforces a particular Direct Support FDC and receives its instructions from that FDC.

(c) General Support. Reinforcing to DS. In this role, the battalion receives requests for fire from both the DS FDC to which it is assigned and from the division artillery FDC.

(d) General Support at D/A. In this role, the battalion receives requests for fire only from the division artillery FDC.

(e) General Support. Reinforcing to D/A. In this role, the battalion receives requests for fire from both the D/A FDC and the Group FDC.

(f) General Support at Group. In this role, the battalion receives requests for fire only from the Group FDC.

2.3.11 RED Artillery Force Organization

A RED artillery battalion may be assigned to any one of three echelons:

(a) Regimental Artillery Group

(b) Divisional Artillery Group

(c) Army Artillery Group

2.4 BLUE FDC Events

As mentioned in describing the target tape, five types of BLUE FDC events may be present on the target tape: fire missions, MET messages, ATI reports, surveys, and fire plans. In addition to BLUE FDC events present on the target tape, there are internally generated BLUE FDC events which are of the fire mission, ATI report, and fire plan types. There are also other FDC related events input to AFSM separate from the target tape such as FDC reliability failure periods and FDC movement schedules.

Let us now examine each of these BLUE FDC event types in turn.

2.4.1 BLUE Fire Mission

2.4.1.1 At the FDC

This event is a request for BLUE artillery fire resulting either from a target acquisition on the target tape or from a request for additional fire (RFAF) from another FDC.

The FDC to which the request for fire is directed examines those artillery resources available to it to determine how much, if any, fire it can contribute.

For each fire mission considered by an FDC, there is an associated level of effects called the required effects. For original fire missions, the required effects level is a function of the input damage levels desired against various target types and the estimated target type. (For example, if the inputs say to shoot for a 10% damage level against tank targets and the estimated target type is a tank platoon, then the required effects level would be 0.10). For RFAF's, the required effects level is the additional level of damage the receiving FDC needs to obtain against the target in order to bring the total effects up to the input damage level for that target type.

The tactical assignment of the FDC will determine the order in which it examines its battalions in trying to mass enough fire to achieve the required effects (See Fig. 1).

The FDC will mass its batteries in order until (1) the required effects level is met, (2) the massing limit is reached, or (3) all of its battalions have been checked, (See Fig. 2).

FIGURE 1 BATTALION PRIORITY AT EACH FDC

FDC

<u>DS</u>	<u>D/A</u>	<u>GROUP</u>
DS Bn	GS (at D/A) Bns	GS (at Group) Bn
then	then	then
R Bn	GSR (to D/A) Bns	GSR (to D/A) Bns
then	then	then
GSR (to DS) Bn	GSR (to DS) Bns	generate RFAF
then	then	to D/A.
generate RFAF to D/A.	DS + R Bns	
	then	
	generate RFAF to Group	

where

DS Bn	means Direct Support Battalion
R Bn	means Reinforcing Battalion
GSR (to DS) Bn	means General Support-Reinforcing Battalion to Direct Support
GS (at D/A)	means General Support at Division Artillery FDC
GSR (to D/A)	means General Support-Reinforcing to Division Artillery FDC
GS (to Group)	means General Support at Group Artillery FDC
D/A	means Division Artillery
RFAF	means Request for Additional Fire

If sufficient additional effects are still required, the massing limit has not yet been reached, and there are battalions at another FDC which have not yet been considered for this fire mission, then an RFAF is generated to another FDC (which FDC depends on the tactical assignment of the sending FDC and the past history of RFAF's of the mission.) This RFAF is added to the list of BLUE FDC events in time sequence (with a delay added to account for transmission time) and is treated as a fire mission.

2.4.1.2 Within Each Battalion

As the FDC attempts to mass enough fire to achieve the required effects, it must examine its assigned battalions for the effects they can achieve. Each battalion is checked by first ordering its batteries and then checking each battery for the effects it can achieve.

The batteries are put in an order that depends on battery-to-target range and time of availability of the battery to shoot the mission. Batteries that are defeated, suppressed, out-of-range, out-of-ammo, too busy, on the move, or not able to fire the minimum number of tubes are ignored in checking for effects.

For each battery that can fire on the target, the model calculates the effects it could achieve against the target using each of its available conventional (HE, ICM) round types. For each HE or ICM round, the level of effects the battery could achieve is calculated using the estimated effects methodology built into the TACFIRE computer subject to ammo availability constraints (based on basic load, resupply rate, previous firings, and ammo set aside for fire plans), maximum volley constraints (an upper bound on the number of volleys any battery having the given weapon system can fire on any one occasion), and hourly rate of fire constraints.

In calculating the effects a given HE round type can achieve, posture sequencing may be applied if the required effects are against an unwarned personnel target (an estimated posture sequence is supplied with the estimated target description.)

In addition, to prevent the wasting of ammo, the effects achieved by each volley of each available round type are calculated and, if it is found that any volley contributes incremental effects less than some cut-off value, then that

volley (and, of course, all subsequent ones of the same round type) is eliminated from consideration for firing by the battery. It may happen that a round is eliminated from consideration because even its first volley cannot achieve the minimum cut-off effects level.

When it is finally determined what level of effects each round type available to the battery can achieve subject to the constraints and cut-off values mentioned, then the round type to be fired is picked. If there are two or more round types that can achieve the required effects level, then the most weight effective (or cost effective, if cost rather than weight is chosen as the criterion) round type is chosen and only as much of it as is needed to achieve the required effects is ordered fired. If only one round type can achieve the required effects level, then as much of that round as required to achieve the required effects level is ordered fired. Finally, if no round type can achieve the required effects level, then the most effective round type is ordered fired.

A similar process goes on in each battery in the battalion under consideration until either the cumulative effects achieved by the batteries considered exceeds the required effects level or all of the batteries in the battalion (except those unable to contribute fire) have been checked.

If additional effects are still required, the FDC will then check its next battalion (if any) in the same battery-by-battery fashion (but with the required effects adjusted downward to take into account the effects to be achieved by the firing orders given to the previously checked battalion) and continue the process through all its battalions until the required effects are achieved, the massing limit is reached, or it has no more battalions to check (if this last condition is reached first, an RFAF may be sent to another FDC.)

The firing orders issued during the checking of the batteries are BLUE battery events and will be executed in time sequence (the only exceptions to this rule are described under GSRS below.)

2.4.1.3 CLGP Fire Missions

Those targets suitable for CLGP fire are treated just like any other fire missions up to a point. The FDC considers its battalions in order, within each battalion the batteries are ordered as usual, and each battery is checked in turn.

However, as soon as any one battery is found that is capable of firing CLGP, the firing orders are issued to it and no further checking of other batteries or generating of RFAF's to other FDC's is done.

If, however, no battery in any of the battalions assigned to the FDC receiving the CLGP mission can fire CLGP rounds within the CLGP window, then the mission is converted to an ordinary non-CLGP mission at the same FDC. (Note: At the present times only forward observer (FO) acquired targets are played as potential CLGP targets and only the battalions assigned to the DS FDC's would have CLGP rounds. For this reason, no attempt is made to RFAF a CLGP mission to a higher echelon FDC until it is converted to a non-CLGP mission.)

2.4.1.4 GSRS (General Support Rocket System)

AFSM can model the use of GSRS. This is done by including one or more batteries of GSRS in the BLUE force mix. Then whenever a fire mission comes into the FDC to which GSRS is assigned, it is checked to see whether it is eligible to receive GSRS fire. In order to be considered for GSRS fire, a target must be an artillery battery, a rocket or missile battery, or an anti-aircraft artillery battery, have a sufficiently large military worth, have a sufficiently large target radius, and not have been fired on previously as a result of the current acquisition.

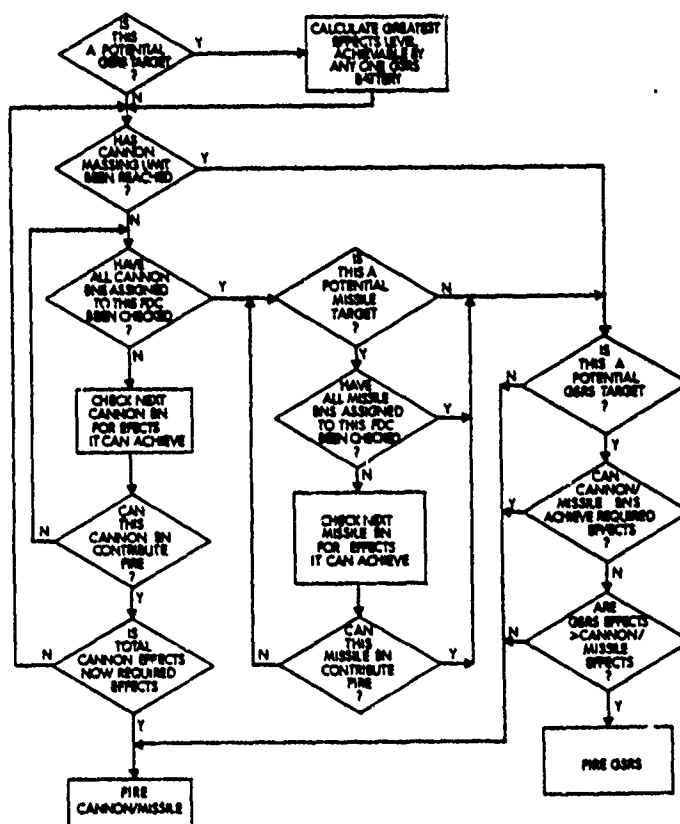
If a target is eligible for GSRS fire, then the effects that GSRS could achieve are calculated and the effects cannon-missile could achieve are calculated (in this case the firing orders to the cannon-missile batteries are not issued at time of checking.)

GSRS is limited to at most one battery volley per target. Cannon-missile are limited by the massing limit on cannon battalions (usually a three battalion limit) and the limit of at most one volley of missiles per target.

Firing orders are then issued to GSRS, if and only if it can achieve greater effects than cannon-missile battalions and cannon-missile battalions cannot achieve the required effects level (See Fig. 2). Otherwise, they are issued to cannon-missile batteries.

GSRS is generally played only at the D/A FDC.

APSM Cannon, Missile, Rocket Massing



NOTE: This is the procedure for massing fires at each higher echelon FDC (D/A and Group).

An AFAP may be generated to another FDC if:

- (a) the cannon massing limit has not been exceeded,
- (b) OGRS and missiles were not selected for firing, and
- (c) the required effects level has not yet been achieved.

The order in which cannon battalions will be massed at each FDC is determined as in Figure 1.

FIGURE 2

2.4.1.5 Missiles

Guided Missiles (such as Lance can be played in AFSM. They are treated the same as cannons except that (1) they are fired only at high military worth targets with sufficiently large target radii, (2) only one battery volley is allowed against any one target, (3) they are fired only against targets where cannon fire has been unable to achieve at least half of the required effects, and (4) they are always added last (i.e., after all cannon battalions at a given FDC have been considered for massing, the missiles may be considered.)

Missiles are generally played only at the Group FDC.

Once missiles have been fired at a target, no further massing is done for that acquisition.

2.4.2 MET Messages

MET messages are read from the target tape and merely keep an FDC busy for a brief period while the MET data is loaded into the TACFIRE.

2.4.3 ATI Reports

In addition to the ATI reports on the target tape itself, there are ATI reports generated within AFSM. When a fire mission (either an original fire mission from the target tape or an RFAF) cannot be processed before the estimated departure time of the target, the effort to mass fire is halted and fire missions against it that have not yet been fired are converted to ATI's.

An ATI merely uses a brief bit of FDC time and has no other effect in AFSM.

2.4.4 Survey

As mentioned above a survey mission is merely a load on the FDC computer.

Note that when mini-moves are played, no new survey missions are generated. This is because of the assumption that a policy of frequent mini-moves would be used only if the battery has PADS, in which case survey is unnecessary.

2.4.5 Fire Plans

A fire plan is a list of areas on the battlefield associated with each of which are both a time (or time interval) at which fire is desired and a volume of fire desired in terms of 155mm HE rounds.

The fire plan message is read from the target tape and directed to a given FDC. That FDC will then generate fire plans for its assigned battalions and schedule the firing of the fire plan missions so as to achieve as much as possible of the fire plan.

The fire plan message is read from the target tape about 45 minutes before the first firing time on the plan. This allows the FDC to schedule the Fire Plan target missions well in advance of the firing time and to do most of the required processing during slack periods between other missions; however, if a firing time is very near and the required fire plan processing is not completed, then the fire plan is force processed at the expense of other missions.

The fire plan processing produces a set of fire plan target missions which are BLUE battery events and which are executed in time sequence. At the time the fire plan target missions are scheduled, the rounds to be fired by each battery on the fire plan are set aside so that they will still be available at the firing time of the fire plan target mission.

2.4.6 Other Blue FDC Events

Blue FDC reliability failures are input (using reliability data on the FDC computer type played - usually TACFIRE, but FADAC can be played) and result in making an FDC go to its lateral back-up (if any) or do its processing by manual methods. In either case, this means the FDC's response time is slower.

BLUE FDC's may move during the game. While an FDC is moving, it processes no missions and its processing is done by its lateral back-up FDC (if any). This also slows down the response time and, of course, puts a greater load on the lateral back-up FDC which is then doing the work of two FDC's.

2.5 BLUE Battery Events

The next major category of AFSM events is that of BLUE battery events. These include executing fire orders from

the battalion (and in turn perhaps from a higher echelon FDC) and executing fire plan target missions.

In addition, BLUE batteries can be suppressed, defeated, or on the move. Being on the move prevents a battery from firing any missions, this usually poses no problem because the FDC knows in advance when a battery is about to move and simply schedules no fire for that battery for that time period. Suppression and defeat are explained in a separate section.

2.5.1 Firing a Fire Mission

When a BLUE battery executes fire orders, it is charged time to fire the mission, the wear on its tubes is updated, its ammo supply is reduced, its effects on the RED target unit are calculated and recorded, and its probability of being acquired by RED counterbattery acquisition devices is increased.

2.5.1.1 HE and ICM

The effectiveness calculations for HE and ICM rounds used to compute the actual damage done to the RED target unit make use of the JMEM Super Quickie II Surface-to-Surface effectiveness methodology (1).

The effects for each battery are obtained from the formula:

$$\bar{f} = ECR * ECD * 1(1 \frac{AEL * NR * REL}{AVP * OF})^{NV * OF} \quad (1)$$

where \bar{f} = fractional casualties (or damage)

ECR = expected fractional coverage of target by weapon pattern in range.

ECD = expected fractional coverage of target by weapon pattern in deflection.

AEL = the single round expected lethal area.

NR = number of rounds per volley.

REL = round reliability.

AVP = volley damage pattern area.

OF = overlap factor.

NV = number of volleys.

The combination of effects from several batteries within the same battalion is done in such a way as to take into account dependency of delivery errors (MPI) within the battalion. The method used is shown in Fig. 3. It tends to overestimate effects slightly though not nearly as much as would result from assuming independence between batteries.

BATTALION EFFECTS FROM BATTERY EFFECTS

Let ECR_I be the expected fractional coverage in range of I_{th} battery's volley(s) on target.

Let ECD_I be the expected fractional coverage in deflection of I_{th} battery's volley(s) on target.

Let P_{K_I} be the probability of kill inside the I_{th} battery's volley area of effects.

Define $ECR' = \text{Maximum } (ECR_1, ECR_2, \dots, ECR_n)$

$ECD' = \text{Maximum } (ECD_1, ECD_2, \dots, ECD_n)$

where $I = 1, 2, \dots, n$

Define

$$P_{K_I}' = \frac{ECR_I * ECD_I * P_{K_I}}{ECR' * ECD'}$$

Then we estimate battalion effects by

$$\hat{f} = ECR' * ECD' * \left(1 - \prod_{I=1}^n (1 - P_{K_I}')\right)$$

Figure 3

2.5.1.2 CLGP

In the case of executing CLGP fire missions, the BLUE battery's effects are a function solely of the number of CLGP rounds fired and the number and types of vehicles (tanks, APC's, trucks) in the target. The number of CLGP rounds fired depends on the number of rounds on hand and the duration of the CLGP window. Once the number of CLGP rounds and number and types of vehicles in the target are known, a table is consulted to obtain the expected number of each vehicle type killed. These values are then converted to fractional damage values for each vehicle type in the target.

2.5.1.3 Updating Damage to RED Units

The fractional damage done to the RED target unit is applied only to that portion of the RED target unit that has survived previous BLUE artillery attacks as well as BLUE nonartillery damage. The updated target unit status is then recorded with the new damage added.

It may happen that the target unit moves before the artillery fire arrives (especially in the case of non-observed fires) in which case the target suffers no damage from the fire.

2.5.1.4 RED Counterbattery Target Acquisition

Each volley fired by a BLUE artillery battery is checked to see whether it leads to RED's acquiring the battery as a potential counterbattery target.

Radar

For each BLUE artillery volley fired, a check is made of each RED counterbattery or countermortar radar in the scenario to determine whether it could track the volley. For each such radar that is turned on, is not defeated, is not suppressed, is not out of commission due to reliability failure, and has coverage of the area from which the volley is fired, the probability of acquisition is calculated as a function of radar-to-battery range, number of other volleys the radar is then trying to track, type of round being tracked, and type of radar. This probability is then compared to a pseudorandom number to determine whether the particular radar being checked acquires the BLUE battery as a result of the volley fired.

Sound

The sound acquisition model is not always played because it requires additional inputs not always available. However, if a record of the sound events from the external wargame that would load a sound system is available and if it is possible to delete from this record those sound events that are replayed in AFSM (BLUE artillery fires and RED counterbattery artillery fires), then it is possible to simulate a sound acquisition system.

Each time a BLUE battery fires a volley, two new sound events are generated (firing and rounds detonating.) The model then looks at the density of sound events from the external wargame record at the current game time as well as the record of AFSM created sound events (from both RED and BLUE artillery firings.) From this information, it estimates the number of sound events that would be occurring simultaneously with the BLUE battery's firing. A probability of acquisition by sound is then found by consulting a table giving probability of acquisition as a function of the number of simultaneous sound events.

This probability is then compared to a pseudorandom number to determine whether the sound acquisition system acquired the BLUE battery.

Moving Target Intelligence

An effort is in progress to add MTI acquisition systems to AFSM. The probability of acquisition of a moving BLUE battery will be a function of the battery's distance behind the FEBA, distance or duration of move, and density of other moves at that game time.

2.5.1.5 Scheduling RED Counterbattery Fire

Once a BLUE battery is acquired by a RED target acquisition device, the resulting acquisition (and its associated TLE which is a function of acquisition range and device type) is reported to the RED counterbattery fire scheduling center. (To prevent overloading the RED CB fire scheduling process, an acquisition unit will not report more than one acquisition of a given BLUE artillery battery site per fifteen minute period.)

If the BLUE battery is already scheduled to receive CB fire at the site reported (as the result of earlier acquisition by the same or other devices), then no further fire

is scheduled against it at this time. Otherwise, the scheduling of RED counterbattery fire now begins.

The RED batteries dedicated to CB fire are examined. Those that are suppressed, defeated, out-of-ammo, out-of-range, too busy, on the move, or unable to fire the minimum number of tubes are eliminated from further consideration. The remaining batteries are considered in an order that depends on the echelon of each battery's battalion, the weapon system of the battery, the range to the target, and the time at which each battery would be available to fire the mission.

RED will then try massing enough batteries to shoot the number of rounds required for neutralization according to RED doctrine. If not enough batteries are able to mass to shoot at least 40% of the required rounds, then no fire orders are issued. Otherwise, fire orders are issued to batteries to fire 40%, 60%, 80%, or 100% (the greatest possible of these alternatives) of the required rounds.

Once the batteries have been picked to fire the RED CB mission, the fire is scheduled to occur time-on-target (TOT) at the time at which the last battery massed can fire. This scheduled fire becomes a RED counterbattery fire event and is executed in proper game time sequence.

2.5.2 Firing a Fire Plan Target Mission

When a BLUE battery executes a fire plan target mission, it is charged time to fire the mission, the wear on its tubes is updated, its ammo set aside by the fire plan for this mission is used, the damage done to the target is calculated, the probability of being acquired by RED counterbattery acquisition devices is increased, and the fire plan score is updated.

These processes are mostly the same as for a BLUE battery executing fire orders as described in 2.5.1. The only major differences are (1) the effectiveness calculations use an older methodology called the K2C method to calculate fractional casualties and damage (if any) and (2) a fire plan score is kept based on the percent of the fire plan rounds successfully delivered at the specified times. This latter fire plan score, is a way of measuring the performance of fire plan execution. (Just counting casualties alone on a fire plan may be misleading since fire plans are often conducted against suspected enemy positions some of which turn out to be unoccupied; so a fire plan could theoretically

be 100% successfully completed by the artillery and produce no effects in terms of casualties or damage.) Of course, the record of casualties and damage is also updated to include the contribution from fire plans.

2.6 RED Counterbattery Fire Event

When the game time reaches the time at which a RED counterbattery fire mission is to be fired, the RED batteries previously scheduled to fire the mission are checked and any that are undefeated and unsuppressed now fire their scheduled quantity of ammunition. (If a battery has lost tubes since the original scheduling of the CB fire, the number of rounds it fires is reduced proportionately).

If the BLUE target battery has moved since the fire was scheduled, there is no damage done. Otherwise, the effects of the RED counterbattery fire on the BLUE battery are calculated and recorded. The effectiveness methodology uses the JMEM Super Quickie II Model with the volley patterns chosen to conform to the RED aiming policy against targets the size of a BLUE artillery battery.

The RED batteries are charged for time spent firing the mission, are charged for ammo used, and have their tube wear updated.

2.7 RED Electronic Warfare

The scenario may designate certain times as periods of RED EW activity. Any FDC transmissions occurring during such a period have a certain nonzero probability of being jammed and hence requiring retransmission. The net effect of this is to worsen (increase) BLUE response time.

Future work is intended to make EW activity also affect communications between target acquisition units and FDC and also to allow EW activity to be restricted to certain sub-sectors of the battlefield.

2.8 Defeat

Each RED unit has a key element. For tank units the key element is tanks, for a mechanized infantry unit the key element is APC's, for a logistics unit the key element may be trucks, for all other units the key element is personnel. In addition, each unit type has a defeat level associated with it. Whenever the total losses of a unit's key element exceed the unit's defeat level, the unit is considered

defeated and further acquisitions of the unit on the target tape are ignored. (This is the one exception to the non-dynamic nature of the target tape.)

For RED counterbattery or countermortar radar units, the unit is considered defeated if either the personnel losses exceed the defeat level or the radar itself is destroyed.

For artillery (both RED and BLUE), a battery is considered permanently defeated if and only its personnel losses exceed the defeat level for its unit type.

A battery that has not been defeated may still be put out of action due to tube losses. This occurs when the number of functioning tubes (or launchers) in a battery drops below the minimum number required for a battery with that weapon system to fire. This drop may come about from attrition, reliability failures, or tube changes. In such a case, the battery is not allowed to take on any new missions; but, because it may in the future get tubes back from the repair shop or get a float tube from its battalion and because it still has sufficient personnel survivors, it may be able to take on missions again at some future game time. Therefore, it is not considered defeated.

2.9 Suppression

Suppression is played explicitly only for artillery batteries (both BLUE and RED) and for RED counterbattery-countermortar radar units. (Suppression of maneuver units as played in the external wargame will, of course, affect their rates of advance or retreat. This, in turn, will have an effect on when and where the maneuver units are acquired which influences the composition of the target tape.)

Each time a battery (RED or BLUE) is being considered to fire a mission, a check is made of its suppression status. This check is made by calculating (for each recent incoming fire) the intensity of fire received in terms of 105mm HE rounds per hectare per minute. (A recent incoming fire is defined to be one that occurred within the last x minutes where x is an AFSM input variable.)

If the intensity of the recent incoming fire is sufficient (according to the USACDC/RARDE suppression model equation) to reduce the level of unsuppressed, surviving personnel below a certain threshold fraction of the original personnel (it nearly always is), then the battery is judged to be

suppressed and unable to fire. Otherwise, it is considered unsuppressed. For armored self-propelled batteries, the personnel inside the howitzers are considered unsuppressed. (Note: The value of x is based on the advice of AMSAA's Tactical Operations Analysis Office and COL Shefi of the Israeli Defense Force).

Whenever a RED counterbattery-countermortar radar unit receives any artillery fire, it is suppressed for twenty-five minutes.

2.10 Mini-Moves

BLUE batteries are allowed to make mini-moves (also termed "shoot-and-scoot", "gun-and-run", "fire-and-flee"). Either one, both, or neither of two types of mini-moves can be played.

The first type of mini-move is played by setting the number of incoming fires a BLUE battery will receive at a site before moving. Then any time during the game that a BLUE battery receives the given number (now usually set at one) of incoming fires at a site, it performs a mini-move.

The other type of mini-move is played by setting the number of volleys that a BLUE battery will fire from a site before moving. Then whenever a BLUE battery completes a fire mission that puts its total volleys from a site over the given number, it performs a mini-move.

In both cases, the mini-move is a short move of up to one kilometer that makes the battery unavailable for a brief period, but generally enhances its survivability by making it more difficult for RED counterbattery fire to catch the BLUE battery at its current site.

Because the movement of RED batteries must agree with the acquired sites of RED batteries on the target tape, it is not possible to play mini-moves for RED batteries in AFSM. A future generation of an AFSM type model may correct this problem.

2.11 Tube Losses and Repair

Both BLUE and RED batteries lose tubes to attrition (enemy fire) and RAM (reliability failures.) Attrition losses are based on the damage done by enemy fire as calculated in the model. RAM losses occur whenever the number of rounds fired (or distance travelled) by the battery since

its last failure exceeds the mean rounds between failures (or mean kilometers travelled between failures.)

Tube losses (of either type) are divided into three categories: short term, long term, and permanent. Short term and long term losses are scheduled for repair and returned to the game when the repair time is up. Permanent losses are unrepairable (at least in the length of the game) and are eliminated from the game; however, the first permanent loss in each battalion is replaced by a float tube.

In addition, tubes may be removed when tube wear exceeds the tube life. In this case, the weapon is out of action for the time it takes to change the tube.

This process of updating each battery tube status (removing tube losses and returning repaired tubes to the battle) is done at regular fifteen minute intervals throughout the game.

2.12 AFSM Outputs

At the end of each game hour AFSM prints out results through that hour.

2.12.1 (1) BLUE artillery force performance measures:

- (a) Military worth of damage to RED target units.
- (b) RED personnel killed by BLUE artillery.
- (c) RED tanks killed by BLUE artillery.
- (d) RED APC's killed by BLUE artillery.
- (e) RED trucks killed by BLUE artillery.
- (f) RED artillery tubes killed by BLUE artillery.
- (g) RED radar killed by BLUE artillery.
- (h) RED A.A. launchers killed by BLUE artillery.

2.12.2 (2) BLUE round expenditures:

- (a) Number of rounds of each type fired by each BLUE battalion.

(b) Cost of rounds of each type fired by each BLUE battalion.

(c) Weight of rounds of each type fired by each BLUE battalion.

2.12.3 (3) Cumulative time during game that each BLUE FDC (battery) was busy processing (firing) artillery missions.

2.12.4 (4) Range tables giving breakdown of round types fired by BLUE at various ranges.

2.12.5 (5) Unaccomplished BLUE artillery mission counter's giving numbers of missions not done for various reasons (out-of-ammo, out-of-range, too busy, etc.)

2.12.6 (6) RAM and attrition results for each BLUE battery gives number of incoming fires received, number of mini-moves made, number of tubes lost due to attrition, number of tubes lost due to RAM, number of tubes currently up, and percent of battery per-sonnel currently surviving.

2.12.7 (7) System performance table breaks down the measures of effectiveness and measures of effort by weapons systems (for BLUE only).

2.12.8 (8) A GSRS table gives data on the performance of the General Support Rocket System in terms of effectiveness, rounds used, etc.

2.12.9 (9) RED Fire unit status tables gives (for each RED battery) number of tubes out due to attrition, number out due to RAM, number currently up, percent of personnel currently alive, number of rounds fired on counterbattery missions.

2.12.10 (1) RED radar unit status gives current damage to each RED counterbattery/countermortar radar.

3. SAMPLE RESULTS

3.1 Mix Descriptions

Because of the classified nature of many of the AFSM inputs, only a very general description is given.

Six cases were run using the same RED attacking threat and the same BLUE defending force. The only items varied

were the types of ammunition available to the BLUE force and the BLUE force movement policy.

The following table summarizes the cases:

<u>Case No.</u>	<u>HE rounds for BLUE?</u>	<u>ICM rounds for BLUE?</u>	<u>CLGP for BLUE?</u>	<u>Mini-Moves</u>
1	Yes	No	No	No
2	Yes	Yes	No	No
3	Yes	Yes	Yes	No
4	Yes	No	No	Yes
5	Yes	Yes	No	Yes
6	Yes	Yes	Yes	Yes

HE rounds were available in all cases and included both RAP (rocket assisted projectile) as well as conventional HE. In cases 2, 3, 5, and 6, the BLUE force also had dual purpose ICM rounds available. In cases 3 and 6, the BLUE force had CLGP available. In cases 1, 2, and 3 no mini-moves were made; in cases 4, 5, and 6 a mini-move was made every time a BLUE battery received fire at a site and every time a BLUE battery fired more than thirty volleys from a site.

3.2 Results

3.2.1 BLUE Force Measures of Effectiveness

Figure 4 shows the relative performance in each case in normalized form (again to avoid possible classification problems.) As one would expect, BLUE force performance improves when ICM is added and a further improvement occurs when CLGP is added (except in number of RED artillery tubes killed obviously CLGP missions against RED armor are diverting a few battery fire missions away from CB missions in the no mini-move cases.)

Each case with mini-moves is superior to the corresponding case with the same ammo-mix but no mini-move policy. This is not surprising since the use of mini-moves increases battery survivability enough to more than overcome the disadvantage of having batteries frequently unavailable.

Looking to Figure 5 we see that in terms of BLUE measures of effort, the differences among ammo-mixes are relatively small. However, the differences between the no mini-move cases and the mini-move cases, clearly show an improvement (when going to mini-moves) in terms of battery survivability

(especially at 8 hours into the game) and in the number of rounds the batteries were able to fire (of course, more rounds fired and better survivability result in more tubes lost to reliability failures).

FIGURE 4 BLUE FORCE MEASURES OF EFFECTIVENESS
RELATIVE CASE PERFORMANCE (BASELINE CASE = 1.00)

CASE	RED LOSSES DUE TO BLUE ARTILLERY			BLUE MISSIONS ACCOMPLISHED
	<u>PERSONNEL</u>	<u>ARMOR[†]</u>	<u>ARTILLERY TUBES</u>	
<u>NO MINI-MOVES</u>				
HE Only (Baseline)	1.00	1.00	1.00	1.00
HE and ICM	1.47	4.36	1.45	1.10
HE, ICM, and CLGP	2.46	30.46	1.40	1.40
<u>MINI-MOVES</u>				
HE Only	1.10	1.09	1.10	1.05
HE and ICM	1.60	5.00	1.75	1.29
HE, ICM, and CLGP	2.81	35.27	1.78	1.70

[†]Armor includes both tanks and APC's.

FIGURE 5 BLUE FORCE MEASURES OF EFFORT
RELATIVE CASE PERFORMANCE (BASELINE CASE = 1.00)

CASE	BLUE TUBES OUT		BLUE ARTILLERY		BLUE ROUNDS		BLUE BATTERIES DEFEATED	
	NO	MINI-MOVES	TOTAL	ATTRITED	RELIABILITY	PERSONNEL LOSSES	FIRE	8 HRS 24 HRS
HE Only (Baseline)	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
HE and ICM	1.06	.97	1.00	1.19	.96	.94	1.08	.94
HE, ICM, and CLGP	1.00	.94	1.00	1.10	1.00	.95	1.00	.94
MINI-MOVES								
HE Only	1.02	.88	1.71	.90	1.13	.17	.94	
HE and ICM	1.26	.84	1.90	.93	1.24	.25	.71	
HE, ICM, and CLGP	1.23	.75	1.95	.92	1.24	.17	.82	

Simplified Diagram of AFSM

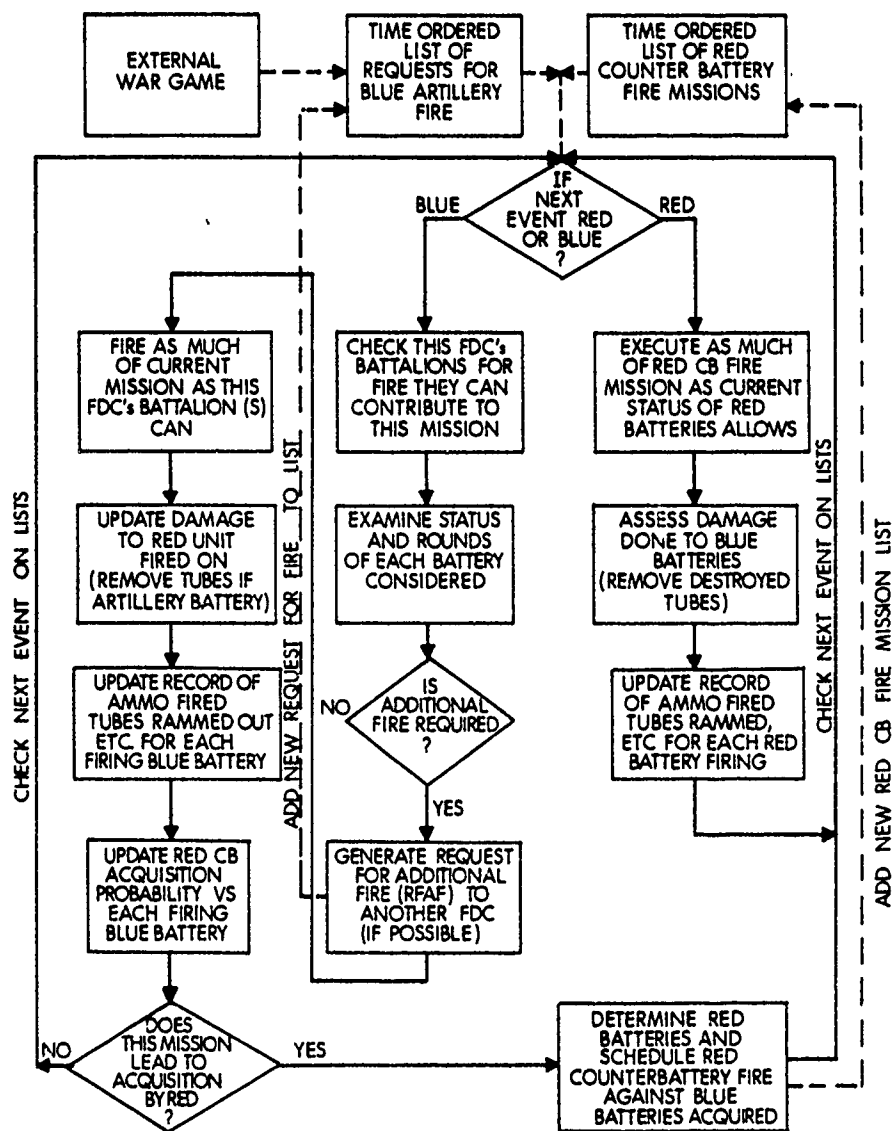


FIGURE 6

REFERENCES

1. Programmable Calculator Manual for Evaluating Effectiveness of Non-Nuclear Surface-To-Surface Indirect-Fire Weapons Against Area Targets, Joint Technical Coordinating Group for Munitions Effectiveness, 61 JTTCG/ME-77-14. This publication documents the JMEM Super Quickie II effectiveness model which is the heart of the AFSM artillery effectiveness methodology.
2. Artillery Force Simulation Model User Manual. This document is still in preparation by contractor, but when completed will give more detail on some aspects of AFSM.

GLOSSARY OF ACRONYMS

AFSM	Artillery Force Simulation Model
ATI	Artillery Target Intelligence
CB	Counterbattery
CLGP	Cannon Launched Guided Projectile
D/A	Division Artillery
DS	Direct Support
EW	Electronic Warfare
FDC	Fire Direction Center
GS	General Support
GSR	General Support Rocket System
GSRS	General Support Rocket System
HE	High Explosive
ICM	Improved Conventional Munitions
MET	Meteorological
MPI	Mean Point of Impact
MTI	Moving Target Intelligence
PADS	Position-Azimuth Determination System
TLE	Target Location Error
TOC	Tactical Operations Center
TOT	Time on Target

THE AMSWAG LIMITED

VISIBILITY STUDY

(LVS)

John J. McCarthy

Frederick M. Campbell

Special Projects Branch

Ground Warfare Division

US Army Materiel Systems Analysis Activity

Aberdeen Proving Ground, MD

ABSTRACT. Over the past two to three years, AMSAA has been involved in studying the influence of limited visibility on combined arms operations. This paper provides a general description of the simulation employed (AMSWAG) and the results of the most recent effort (LVS).

THE AMSWAG LIMITED VISIBILITY STUDY (LVS)

1. INTRODUCTION

In connection with AMSAA's recent involvement in a study of the effects of limited visibility on combined arms operations, the AMSWAG combat simulation was run for a series of cases which addressed the situation in question. This paper presents an overview of the AMSWAG process and a brief summary of the simulation results.

2. AMSWAG OVERVIEW

As indicated in Figure 1, AMSWAG is a time sequenced, battalion level, combined arms combat simulation which is based on Lanchester equations and produces Expected Value (Deterministic) results.

The overall process (Figure 2) consists of a series of preprocessor results which are fed to the main model which produces time sequenced results for subsequent analysis.

The basic system data, which consist of weapon/round performance and system vulnerability characteristics are processed and stored on disc files for subsequent access. These data are primarily study-independent and are constantly being expanded and/or updated.

The study dependent information normally consists of the scenario (a description of the combat conditions and force structure), the terrain (a digitized description of the combat location with respect to an x, y, z, vegetation properties), the terrain/vehicle mobility dependency, and the preselected deployment of the defender locations and the attacker routes of advance. These study-dependent data are processed such that line-of-sight (LOS) as a function of exposure (hull defilade or fully exposed), individual system position and velocity, the presence of mines and/or obscuration are available for each 10 second game interval.

3. REPRESENTING LIMITED VISIBILITY

Approximately three years ago, AMSAA entered into the area of representing the influence of reduced visibility on combat simulation results. Our first efforts were limited to reduced "opening-range" cases, then we came into contact with the Night Vision Laboratories at FT Belvoir, Virginia and used a table look-up data base in conjunction with model

AMSWAG CHARACTERISTICS

COMBINED ARMS COMBAT SIMULATION
BATTALION LEVEL
EXPECTED VALUE (DETERMINISTIC)
LANCHESTER THEORY
TIME SEQUENCED (10 SECONDS)
HIGH RESOLUTION (INDIVIDUAL UNIT)

FIGURE 1

OVERVIEW OF AMSWAG PROCESS

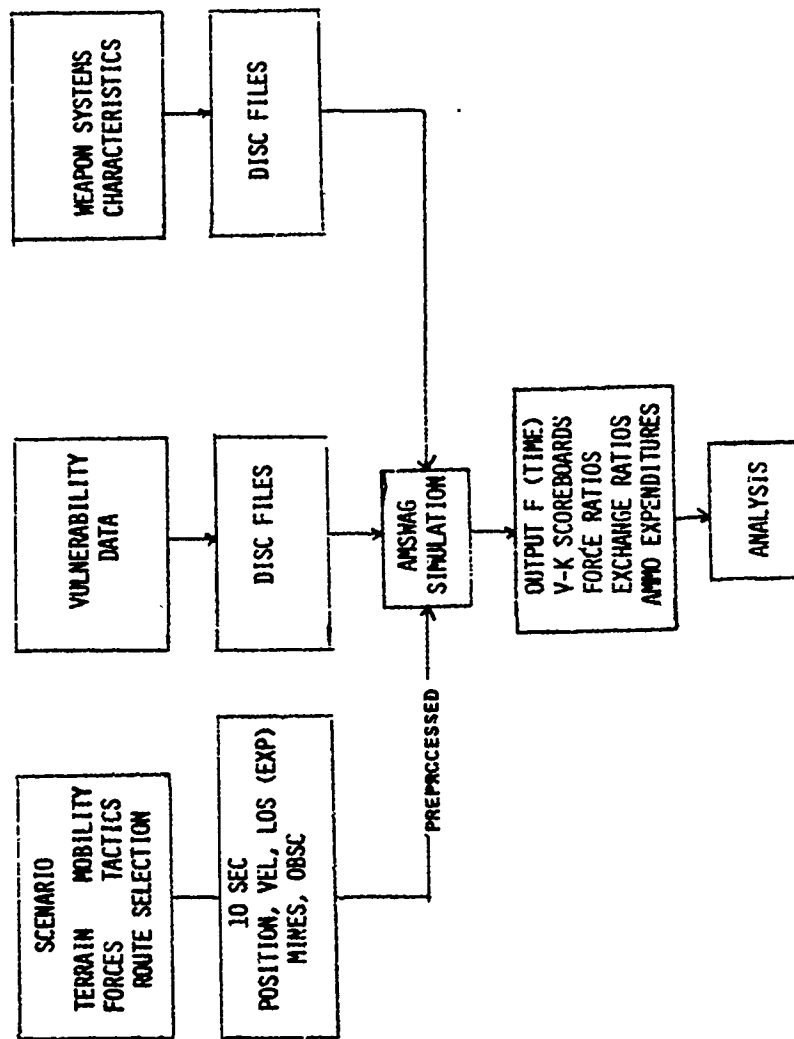


FIGURE 2

changes (Night Vision Net Technical Assessment Study). A joint effort between NVL, AMSAA and CACDA (Combined Arms Combat Development Activity, FT Leavenworth, Kansas) resulted in an algorithm for the large scale NVL model. This model was employed in the Battlefield Illumination Study. The acquisition process which was employed in this study is highlighted on Figures 3 and 4.

Target acquisition is described as a function of the total ability to detect targets, P_{∞} , and an exponential function which represents the search process. The variable T is accumulative as long as LOS is continuous; the process is bypassed if LOS does not exist.

The two primary parameters in the acquisition process, P_{∞} and \bar{t} , are influenced by many parameters in the algorithm. The main variables are: visibility range, device response wavelength, field of view, magnification, target size and contrast, range, and atmospheric attenuation.

4. THE LVS EFFORT

As shown on Figure 5, the LVS considered a battalion attacking force of 30 tanks and 10 APC's versus a reinforced tank company consisting of 10 tanks, 3 long and 3 short range missile systems. A weighting scheme was employed and is displayed on the viewgraph.

The measure of effectiveness, Figure 6, for this study was rather unique, at least with respect to previous AMSWAG efforts. A base case was defined and the MOE was the Normalized Parity Force Multiplier (NPFM).

Several cases were run for each situation; the defender was held at a constant level; however, the attacker force was uniformly increased for each run. The Attacker Force Multiplier, AFM, is the number of 2.66 to 1 forces which was involved. The Parity Force Multiplier, PFM, is the AFM which resulted in 50 percent losses for both sides. The NPFM is then the PFM for each case divided by the PFM for the base case.

The case matrix, displayed on Figure 7, for the LVS effort addressed several visibility conditions (8, 4, 2, and 1 km), day (100 foot candles) and night (.0001 foot candles), as well as both forces using optical, image intensifier and thermal viewing devices.

The Blue Optical/Red Optical, BORO, Day results are shown on Figure 8 in order to display the analysis process. The

TARGET ACQUISITION

$$P_D(T) = P_{(\infty)} \left(1.0 - e^{-\frac{T}{\bar{T}}} \right)$$

WHERE:

$P_D(T)$ IS THE PROBABILITY OF DETECTION AT ANY
GIVEN TIME, T

$P_{(\infty)}$ IS THE PROBABILITY OF DETECTION GIVEN AN
INFINITE SEARCH

\bar{T} IS THE MEAN TIME OF THE SEARCH AT A GIVEN
RANGE

FIGURE 3

LVS IMPACT

ACQUISITION:

P_{∞} , \bar{T} , PPM (NEAR 0 DUE TO APPLICATION & TRIAL RESULTS)
VISIBILITY RANGE (2% TRANSMISSION OF 100% CONTRAST TARGET)

DEVICE:

FOV
MAG
WAVE LENGTH (VISUAL OR THERMAL)

RANGE:

ATTENUATION (ATMOS)
TARGET SIZE
CONTRAST (VISUAL = .25, $\Delta T \approx 2^{\circ} C$)
SKY/GROUND RATIO

MEMORY (IF LOS IS CONTINUOUS T GROWS)

FIGURE 4

FORCES 1985

<u>DEFENDER</u>	<u>FORCE VALUE</u>
10 TANKS	100
3 ATGM (LONG RANGE)	24
3 ATGM (SHORT RANGE)	<u>15</u>
	139
 <u>ATTACKER</u>	
30 TANKS	300
10 APC (IMPROVED ATGM)	<u>70</u>
	370
 BASE INITIAL FORCE RATIO =	$\frac{370}{139} = 2.66$

FIGURE 5

MEASURE OF EFFECTIVENESS

ATTACKER FORCE MULTIPLIER (UNIFORM INCREASE FOR EVERY WEAPON UNIT)

AFM = 1 IMPLIES 2.66 INITIAL FORCE RATIO

END-OF-GAME = 50% ATTACKER LOSSES OR ATTACKER REACHES OBJECTIVE
(500M FROM DEFENDER POSITIONS)

FIND AFM SUCH THAT END-OF-GAME IS 50/50, I.E., PARITY.

PARITY FORCE MULTIPLIER = AFM WHICH GIVES 50/50

NORMALIZED PARITY FORCE MULTIPLIER = $\frac{\text{PFM GAME}}{\text{PFM BASE}}$

FIGURE 6

AMSWAG CASE MATRIX

VISIBILITY RANGE (KM)			
<u>DAY</u>		<u>NIGHT</u>	
8	Blue	4	Blue
2	Optical	2	Image Intensifier
1.5	Red	1	Red
1	Optical		Image Intensifier
8	Blue	4	Blue
2	Thermal	2	Thermal
1	Red	1	Red
	Thermal		Thermal

FIGURE 7

AMSWAG Individual Case Results (Primary Case Matrix)

Optical Detection Devices (BORO), Day Light Level, No Smoke

<u>Visibility Range (KM)</u>	<u>Attacker Force Multiplier</u>	<u>Percent Losses</u>	
		<u>Attacker</u>	<u>Defender</u>
8 (Base Case)	1.0	50.0	5.9
	2.0	50.0	24.2
	2.5	50.0	50.0

PFM = 2.5

NPFM = 1.0

2	1.0	50.0	5.7
	1.5	50.0	14.8
	2.0	50.0	27.0
	2.5	43.8	68.8

PFM = 2.4

NPFM = 1.0

1.5	0.8	50.0	10.0
	1.0	44.0	12.0
	1.5	32.0	21.0

After approximately 20 minutes of battle, both forces lost 30% with a 1.7 to 1.9 attacker force multiplier,

NPFM = .7

1.0	0.5	16.6	0.8
	1.0	12.5	2.0
	2.0	7.4	5.7

This game ends prior to significant losses, both forces lose 6% after 40 minutes; this implies an NPFM = .9 which should only be used to show trends.

FIGURE 8

AMSWAG Individual Case Results (Primary Case Matrix)

Image Intensifier Devices (BIRI), Night Light Level. No Smoke

<u>Visibility Range (KM)</u>	<u>Attacker Force Multiplier</u>	<u>Percent Losses</u>	
		<u>Attacker</u>	<u>Defender</u>
4	1.0	50.0	2.0
	2.0	50.0	11.8
	2.8	50.0	22.2
	3.0	50.0	31.5
	3.5	50.0	47.0
PFM = 3.6		NPFM = 1.4	
2	1.0	50.0	0.5
	2.0	50.0	3.0
	3.0	50.0	9.5
	3.5	50.0	16.5
PFM = 4.5		NPFM = 1.8	
1	0.5	44.8	0.0
	1.0	30.8	0.0
	2.0	21.8	0.0

Under the conditions modeled, constant light level = .0001 foot candles, the defender can apparently acquire a portion of the attacker as he arrives at his final location and kills that portion; the attacker cannot acquire the defender. There is no force ratio which will yield equal losses and the attacker can reach his final objective.

FIGURE 9

AMSWAG Individual Case Results (Primary Case Matrix)

Thermal Detection Devices (BTRT)*, Day Light Level, No Smoke**

<u>Visibility Range (KM)</u>	<u>Attacker Force Multiplier</u>	<u>Percent Losses</u>	
		<u>Attacker</u>	<u>Defender</u>
8	1.0	50.0	3.1
	2.0	50.0	14.7
	3.0	47.0	67.9
	4.0	29.1	66.3
PFM = 2.9		NPFM = 1.2	
2	1.0	50.0	3.1
	2.0	50.0	13.5
	3.0	49.4	67.8
	4.0	31.2	67.1
PFM = 3.0		NPFM = 1.2	
1	1.0	50.0	4.0
	2.0	50.0	20.6
	2.5	50.0	42.1
	3.0	43.9	67.1
PFM = 2.5		NPFM = 1.0	

*Blue Thermal/Red Thermal

**The cases where the attacker used smoke gave the same results since the devices were capable of seeing through the clouds generated.

FIGURE 10

AMSWAG Individual Case Results (Primary Case Matrix)

Thermal Detection Devices (BTRT)*, Night Light Level, No Smoke

<u>Visibility Range (KM)</u>	<u>Attacker Force Multiplier</u>	<u>Percent Losses</u>	
		<u>Attacker</u>	<u>Defender</u>
4	1.0	50.0	0.0
	2.0	50.0	3.2
	3.0	50.0	11.0
	4.0	50.0	23.0
PFM > 4.0		NPFM 1.6	
2	1.0	50.0	3.1
	1.5	50.0	8.0
	2.0	50.0	12.8
	3.0	50.0	44.0
	4.0	36.8	70.9
PFM = 3.0		NPFM = 1.2	
1	1.0	50.0	5.8
	1.5	50.0	13.0
	2.0	50.0	27.0
	2.5	48.2	67.8
PFM = 2.2		NPFM = .9	

*Blue Thermal/Red Thermal

FIGURE 11

other results are available in Figures 9, 10, and 11. The Blue Image Intensifier/Red Image Intensifier, BIRI, case should be viewed with discretion for the following reasons:

4.1 Light level, as modeled, was maintained at .0001 foot candles.

4.2 Apparently, the light level was such that the defender could just see the attacker and attrite him as he advanced and the attacker never really was able to acquire the defender.

4.3 The 4.1 and 4.2 conditions above are unrealistic in that the battle conditions (firings and burning targets, use of artificial illumination) would have produced more light and consequently the battle would have been much less favorable to the defender.

5. OBSERVATIONS

Figure 12 summarizes the results obtained. Excluding the BIRI cases, the trend is certainly that the defender's situation gets worse as the not uncommon 2 km visibility condition is approached. There does appear to be an increase associated with the 1000 meter situation which suggests the impact of the short range missile systems. The limited visibility situations highly suggest that there will be a greater emphasis on the smaller, more mobile short range systems.

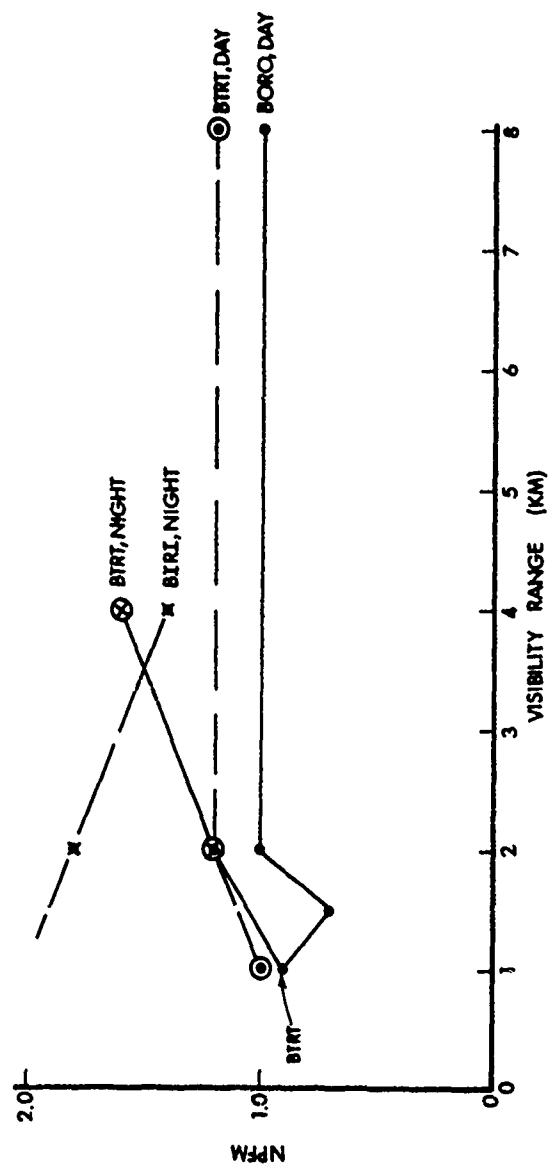
Certainly the more extreme conditions, i.e., less than 1 km visibility, present far greater problems.

There appears to be a great need for training under these conditions.

6. CONTINUING EFFORTS

With respect to our future efforts, the search portion of the acquisition process will certainly receive improvements which will reflect the changing search area as the battle progresses and the effects of the target's motion. These are ongoing projects which should contribute toward this revision.

The current efforts toward modeling the influence of smoke and dust obscuration is in its developmental phase.



Normalized Parity Force Multiplier (NPFM) vs Visibility Range.

FIGURE 12

JOINT MUNITIONS EFFECTIVENESS MANUAL (JMEM) AND
APPLICATIONS OF DATA

John J. McCarthy

US Army Materiel Systems Analysis Activity
Aberdeen Proving Ground, Maryland

ABSTRACT: Joint Munitions Effectiveness Manuals are developed by the Joint Technical Coordinating Group for Munitions Effectiveness (JTTCG/ME). The methodology and data base to develop weapons effectiveness estimates have been standardized and JMEM's developed for many of the US Operational Weapons.

Critical data requirements and methodology considerations will be presented of surface-to-surface, air-to-surface and anti-air weapons. Sensitivity of weapons effectiveness in each of the above categories will be presented. The use of small computers in addressing the weaponizing problem will be presented.

Special studies conducted using the JMEM data base and methodology will be presented. Specific studies will address the use of weapons effectiveness in analyzing tank gunnery and training, artillery fire techniques and support equipment for fire planning.

A brief discussion of target acquisition and its impact on air-to-surface weapons effects and employment will be presented.

JOINT MUNITIONS EFFECTIVENESS MANUAL (JMEM)
AND APPLICATIONS OF DATA

1. INTRODUCTION

Munitions Effectiveness data is the basic input to many military OR/SA studies. In the US, the Joint Technical Coordinating Group for Munitions Effectiveness has the responsibility for standardizing the methodology and input data, as well as developing effectiveness estimates for all conventional weapons. These data are published in the Joint Munitions Effectiveness Manuals (JMEM).

This paper will briefly discuss the basic measures of effectiveness, some critical input data and the sensitivity of weapons effectiveness to critical parameters. In addition, results of a special study on artillery firing techniques will be presented.

2. MEASURES OF EFFECTIVENESS

The most basic and frequently used measures of effectiveness are vulnerable area and lethal area. The term vulnerable area is used for single fragments or weapons attacking targets such as trucks and tanks. Lethal area is generally used to evaluate fragmenting weapons which have effects against many target elements distributed over a large area on the ground.

For targets, such as tanks and trucks, if the presented area from a single aspect angle for an attacking weapon is considered and the target is divided into rectangular cells, then the basic measure of effectiveness, vulnerable area A_v , of the attacking weapon is defined as

$$A_v = \sum_x \sum_y P_K(x,y) \Delta x \Delta y$$

for cells of uniform size where $P_K(x,y)$ is the probability of killing the target about point (x,y) with area $\Delta x \Delta y$ of the cell.

For fragmenting weapons, lethal area is defined similar to vulnerable area except that the projection of fragments over an area is considered against many target elements. Lethal areas is defined as

$$AL = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} P_K(x,y) \, dx dy$$

where $P_K(x,y)$ is the probability that a target element at point (x,y) is incapacitated.

Given a lethal area for a fragmenting munition, the JTCG/ME has developed standard methodologies to generate expected fractional damage (F_D) estimates for given weapon expenditures against a variety of target types. The effectiveness estimates for all conventional weapons are published in Joint Munitions Effectiveness Manuals.

In order to allow Analysts and Planners to estimate the effectiveness of weapons for different delivery accuracies, target, delivery conditions, etc., the JTCG/ME effectiveness methodologies have been adapted to small computers such as the Wang 700, Wang 2200, HP65, and HP67/97. These methods are used widely in training and wargames.

The basic mathematical formula used to compute expected fractional damage for artillery is:

$$F_D = EC_R(EC_D) \left\{ 1.0 - \left[1.0 - \frac{N_R(r_R)(A_{EL})}{A_{VP}(OF)} \right]^{N_V(OF)} \right\}$$

F_D = Expected Fractional Damage

EC_R = Expected Fractional Coverage of the Target by the Weapon Pattern in Range

EC_D = Expected Fractional Coverage of the Target by the Weapon Pattern in Deflection

N_R = Number of Rounds per Volley

r_R = Round Reliability

A_{EL} = Single Round Expected Lethal Area

A_{VP} = Volley Damage Pattern Area

OF = Overlap Factor

N_V = Number of Volleys

The expected coverage of the target area is EC_R times EC_R

and is computed using the total delivery accuracy for the entire weapon system. The delivery accuracy is assumed to be a normal distribution. Range and deflection are computed separately since for most artillery systems, the errors are larger in range than in deflection. Similar methods have been developed for air-to-surface weapons which include unitary warhead, guided warheads, rockets, and cluster weapons.

3. CRITICAL INPUT DATA

A necessary element to determine the effectiveness of a given weapon system is input data. In general, the data required to compute weapons effectiveness can be categorized as:

- Munitions Characteristics
- Delivery Accuracy
- Reliability
- Target Vulnerability
- Battlefield Scenario - number of weapons, target size, range-to-target, etc.

3.1 Munitions Characteristics

Fragmenting munitions, such as bombs and artillery projectiles, are statically detonated in arena tests to collect fragmentation and velocity data. Fragment mass distribution data are collected by recovering sample fragments as a function of the angle off the nose of the shell. This is generally done for an angle of 0 to 180 degrees. Velocity data are collected from 180 to 360 degrees by the use of aluminum panels and high speed cameras.

Figure 1 is an example of the reduced arena data for a typical projectile. Generally, three to five projectiles are tested and results are averaged.

In order to evaluate shaped charge munitions, tests are conducted against armor plate. Shaped charge rounds are tested at different standoff distances and penetration and penetration hole diameter data are collected. Figure 2 shows a typical test set-up and the type data developed.

Penetration data are also collected for kinetic energy projectiles by testing against armor plate. Tests are conducted against various armor plate thicknesses and penetration data are developed as a function of striking velocity. In addition, fragmentation (spall) data are collected behind the armor. The number, mass, velocity and angular distribution of fragments are recorded for tests conducted as shown in Figure 3.

Tests similar to these are conducted with different threats and target materials to allow evaluation of surface-to-surface, air-to-surface and anti-air weapons against a wide range of targets.

3.2 Delivery Accuracy

Delivery accuracy of weapons systems is addressed both analytically and through testing. For artillery and direct fire systems an errors budget approach is used. Figures 4 and 5 show the component errors considered for artillery and tank systems.

In artillery fire we have two basic errors - Precision and Mean Point of Impact (MPI). Precision error is the scatter of burst points about the (MPI) of a group of rounds fired from a single weapon on a single occasion. The Mean Point of Impact is the scatter of the MPI's about an aimpoint (target). The components that make up the MPI error are shown for an unadjusted fire technique called MET & VE. If observer-adjusted fire is used, these component MPI errors are zero and the only MPI error is the error is adjustment. Figure 6 shows typical type data for an artillery system.

In the case of tank gunnery there are three primary errors - fixed bias, variable bias and random error. The components are shown in Figure 5. The variable bias error is a function of the conditions at the time of firing.

Generally accuracy data for all systems are compared with field test results and whenever possible combat data is used. In some cases, especially air-to-surface weapons, tests are conducted under unfamiliar conditions in attempt to better estimate the true accuracy under combat conditions.

In the anti-air area, the problems are more complex and it is difficult to obtain extensive test data. Evaluation of missile systems, both air defense and air-to-air, rely heavily on detail fly-out simulations. These are supplemented by limited test firings.

3.3 Reliability

Reliability data is a critical factor in the overall assessment of a weapons performance against a target. Reliability data are developed for both warhead functioning as well as fuze functioning. If the system being evaluated is more complex, such as a rocket and/or guided systems, the reliability of these elements is also included. These data are generally collected through development tests and are continually updated when tests are conducted against actual targets.

3.4 Target Vulnerability

The final critical input data to the weapon effectiveness problem is target vulnerability. In order to develop target vulnerability estimates a defeat criteria is defined. In the case of armored vehicles, such as tanks, the defeat criteria are:

- Mobility/Fire Power (M/F)
- Catastrophic - Not Repairable (K)

Having defined these criteria, details of the target are examined to determine which components contribute to the kill. Generally, a computer description of the target is developed as shown in Figure 7. A grid as shown in Figure 8 is considered, and for a given threat shot lines are examined as shown in Figure 9 to determine which elements are damaged and whether or not they contribute to a kill. Relating the performance of a threat against armor, and the vulnerability of components, estimates are made of the total system vulnerability.

This general approach to target vulnerability has been used to develop estimates for:

- Trucks
- Tanks
- Aircraft
- Bridges
- Armored Personnel Carriers

and many other targets.

A more generalized form of vulnerability has been developed for personnel targets. Figure 10 shows the form of the equation used to estimate the probability of incapacitation given a hit by a fragment. Also shown are four casualty criteria defined. Constants for each casualty criteria have been developed based on firings into gelatin and correlating these results with expected damage to a human being.

3.5 Battlefield Scenario

Having collected the critical input data on munitions characteristics, delivery accuracy, reliability and target vulnerability it is possible to evaluate a weapon system under a variety of battlefield conditions. In the case of artillery, it is possible to evaluate the sensitivity of weapons effectiveness to:

- Weapon-to-target range
- Target size and type
- Firing formation and technique

In addition, it is possible to examine the sensitivity of the effectiveness to anyone of the critical input data. Figures 11 and 12 show the sensitivity of artillery effectiveness to changes in:

- Precision probable error in range
- Target size
- Firing technique.

Given the input discussed in the paper, it is possible to generate these data using the small computer methods.

4. ENVIRONMENTAL EFFECTS

Environmental considerations can significantly influence the estimates of weapons effectiveness. There are many environmental factors that can or should be considered. However, examples will be limited to:

- Effects of terrain shielding on weapon effectiveness
- Effects of vegetation
- Effects of terrain on target acquisition.

4.1 Effects of Terrain Shielding

The terrain in which a target is attacked significantly effects the effectiveness of fragmenting weapons. Field measurements have been carried out to characterize several types of terrain found in Maryland. Figure 13 shows the lethal area as a function of burst height for typical fragmenting shell and bomb. Results are presented for a rough, rolling and flat terrain.

The results presented in Figure 13 show the sensitivity of the effects to terrain shielding as a function of burst height. This type data can be used to evaluate the potential benefits of airburst versus ground burst fuzes. It also shows the importance of properly considering the terrain.

4.2 Vegetation Effects

During the Vietnam War considerable effort was expended on testing munitions in various vegetation-environments such as marsh grass, temperate and tropical forests. This testing was required because generally, weapons testing and evaluation had been carried out in open terrain and the observed performance in a jungle environment was quite different.

The JTCG/ME conducted many tests as well as field characterization efforts. As a result, we now have standard environments defined and consider them in evaluating all weapons. Figure 14 shows the reduction in weapons effectiveness for artillery projectiles when used in marsh grass or forest environments.

4.3 Effects of Terrain on Target Acquisition

The final area concerning the environment concerns target acquisition. Acquisition of a target is a function of line-of-sight for both air and ground delivered weapons. In the case of air targets, the JTCG/ME has prepared a detailed report on those factors which effect the successful launch of an air delivered weapon. This report addresses the delivery mode, weather, terrain, target contrast, and many other parameters. Figure 15 is an example of the type data presented and shows the sensitivity to variations in the terrain.

5. APPLICATION OF WEAPONS EFFECTIVENESS

The weapons effectiveness data and methodologies have many applications. The OR/SA community uses the data in support of many wargames. In addition, these data can be used to evaluate product improvements, changes in deployment and firing techniques and new systems. An example of such an application is a study done for the artillery community on evaluating alternate firing techniques for consideration in a Battery Computer System (BCS).

Figure 16 shows the various techniques evaluated against a given target. The Fendrikov is a technique found in Soviet Literature, whereas the Converged, and Lazy W are standard. The BCS technique is the firing pattern included in the proposed BCS.

In evaluating the various techniques, it was found that a modified Fendrikov technique is desirable. The specific spacing of aimpoints on a given target is determined by the algorithm shown in Figure 17. The spacing is illustrated in Figure 18. The benefits of using this technique over the standard Lazy W is shown in Figure 19.

6. SUMMARY

Standardization of the weapons effectiveness methodologies and the input data has resulted in extensive OR/SA efforts in the US and other countries. This standardization has allowed study results to be compared with a reasonable confidence that the data base used is consistent. Very little has been presented on anti-air weapons. This is a relatively new effort in the JTCG/ME. Efforts have been underway for sometime to develop standard methodologies and expand the data base. Success in this area should provide a total data base or weapons effectiveness.

EXAMPLE OF REDUCED ARENA DATA

WGT INTERNAL	0-10°		10-20°		20-30°		30-40°		40-50°		ETC.
	m	N	m	N	m	N	m	N	m	N	
1-2	1.4	72	1.6	56	1.5	40	1.4	31	1.3	140	
2-5	3.7	17	4.2	41	2.8	28	4.0	40	3.9	120	
5-10	7.2	21	8.1	73	9.0	68	8.7	90	7.8	211	
10-20	15.1	14	11.2	54	17.1	92	18.2	112	17.9	201	
ETC.											
250-500	303	2	460	3	210	1	—	0	—	0	

ZONAL VELOCITY (FPS)	4750	4980	5100	5490	6320	ETC.
-------------------------	------	------	------	------	------	------

FIGURE 1

PROJECTILE CHARACTERISTICS

● SHAPE CHARGE

- A WARHEAD CHARGE DIAMETER
- B BUILT IN STANDOFF
- C PENETRATION VS STANDOFF
- D PROFILE HOLE DIAMETER VS STANDOFF AND DEPTH OF PENETRATION

● KINETIC ENERGY AP:

- A PROJECTILE DIAMETER
- B PENETRATOR DIAMETER
- C VELOCITY AS A FUNCTION OF RANGE
- D BALLISTIC LIMIT FOR VARIOUS CONDITIONS OF PLATE THICKNESS AND OBLIQUITIES

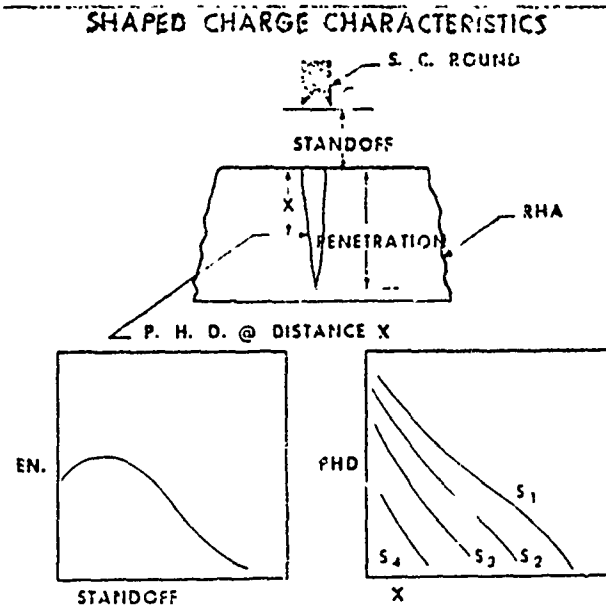
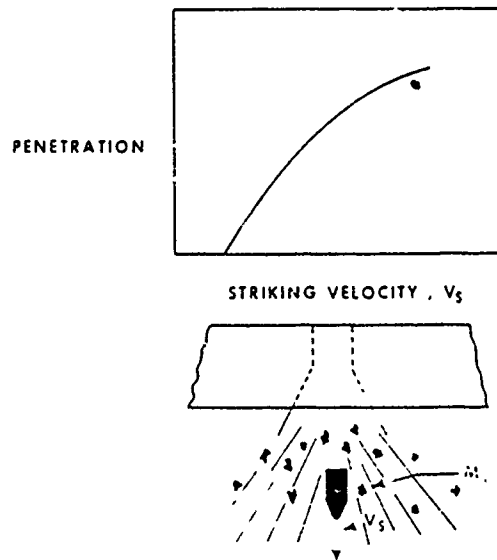


FIGURE 2

AP PROJECTILE CHARACTERISTICS



VULNERABILITY ASPECTS OF PROBLEM

MUST DETERMINE THE PROBABILITY THAT THE
TARGET IS RENDERED INEFFECTIVE BY THE WEAPON
(BLAST EFFECTS; FRAGMENT EFFECTS; SHAPE CHARGE
EFFECTS; OR COMBINATIONS THEREOF)

FIGURE 3

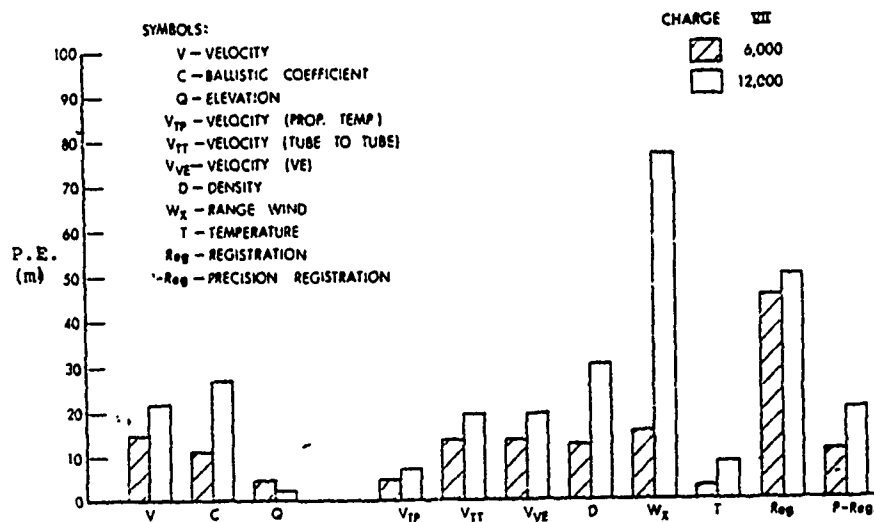
FIGURE 4 ERROR SOURCES

<u>OBSERVER ADJUSTED</u>	<u>MET & VE</u>
Precision	Precision
MPI	MPI
Adjustment	Registration
	VE
	Wind
	Density
	Temperature
	Velocity
	Temperature
	Tube to Tube
	Target Location

ERROR SOURCES

- | I FIXED BIAS | II VARIABLE BIAS | III RANDOM ERROR |
|--------------|-----------------------------|------------------|
| ● PARALLAX | ○ CANT | ○ ROUND-TO-ROUND |
| ● DRIFT | ○ WIND | ○ LAY |
| ● MEAN JUMP | ○ MUZZLE VELOCITY VARIATION | |
| | ○ RANGE ESTIMATION | |
| | ● FIRE CONTROL | |
| | ● AIR TEMPERATURE | |
| | ● AIR DENSITY | |
| | ● OPTICAL PATH BENDING | |
| | ● WINDAGE JUMP | |
| | ● ZEROING | |
| | ● JUMP | |

FIGURE 5



155 mm Howitzer, M109, with Projectile, HE, M107
Artillery Delivery Errors

FIGURE 6

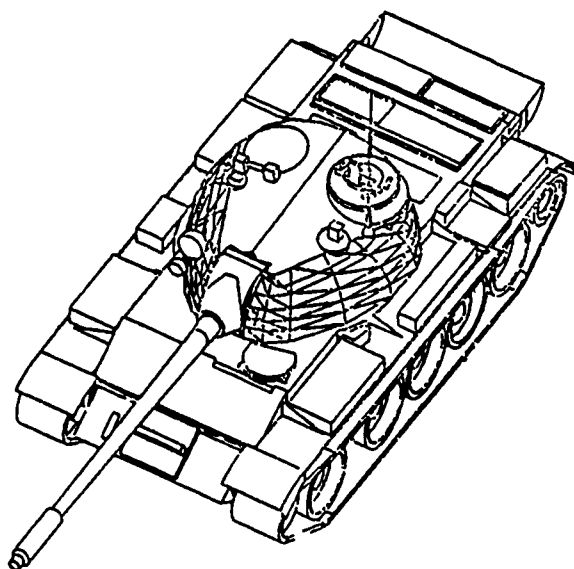
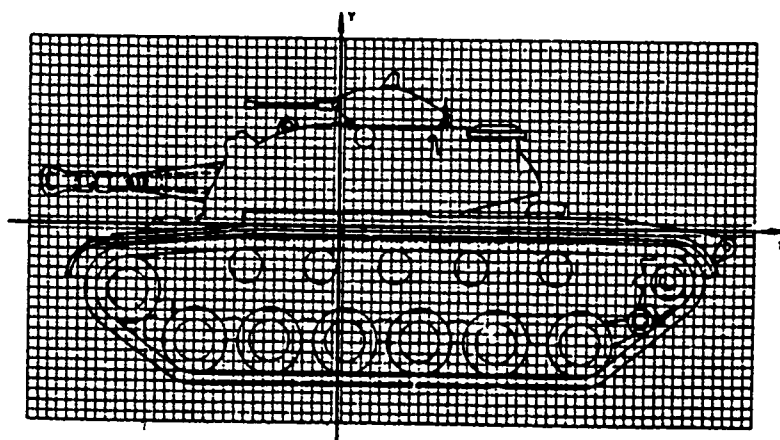
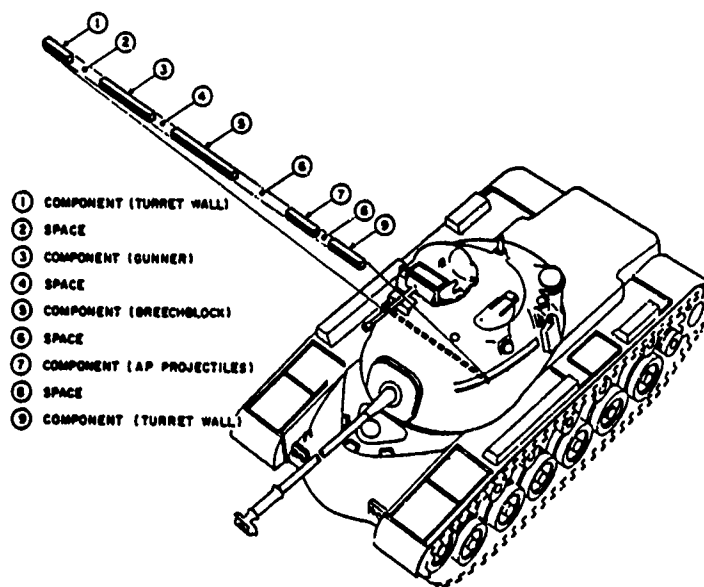


FIGURE 7



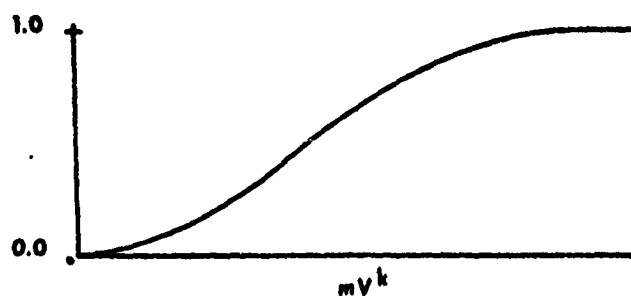
Grid Superimposed on 90° Azimuth View of Target.

FIGURE 8



Representative Vehicle Section for Target Cell Description Data

FIGURE 9



$$P_{I/H} = 1.0 - \exp(-A[mV^k - B]^n)$$

m = WEIGHT

v = VELOCITY

A, k, B, n = CONSTANTS AS A FUNCTION OF CASUALTY CRITERIA

'STANDARD' CRITERIA

- (1) DEFENSE: 30 SECONDS
- (2) ASSAULT: 30 SECONDS
- (3) ASSAULT: 5 MINUTES
- (4) SUPPLY: 12 HOURS

FIGURE 10

EFFECT OF PRECISION ERROR

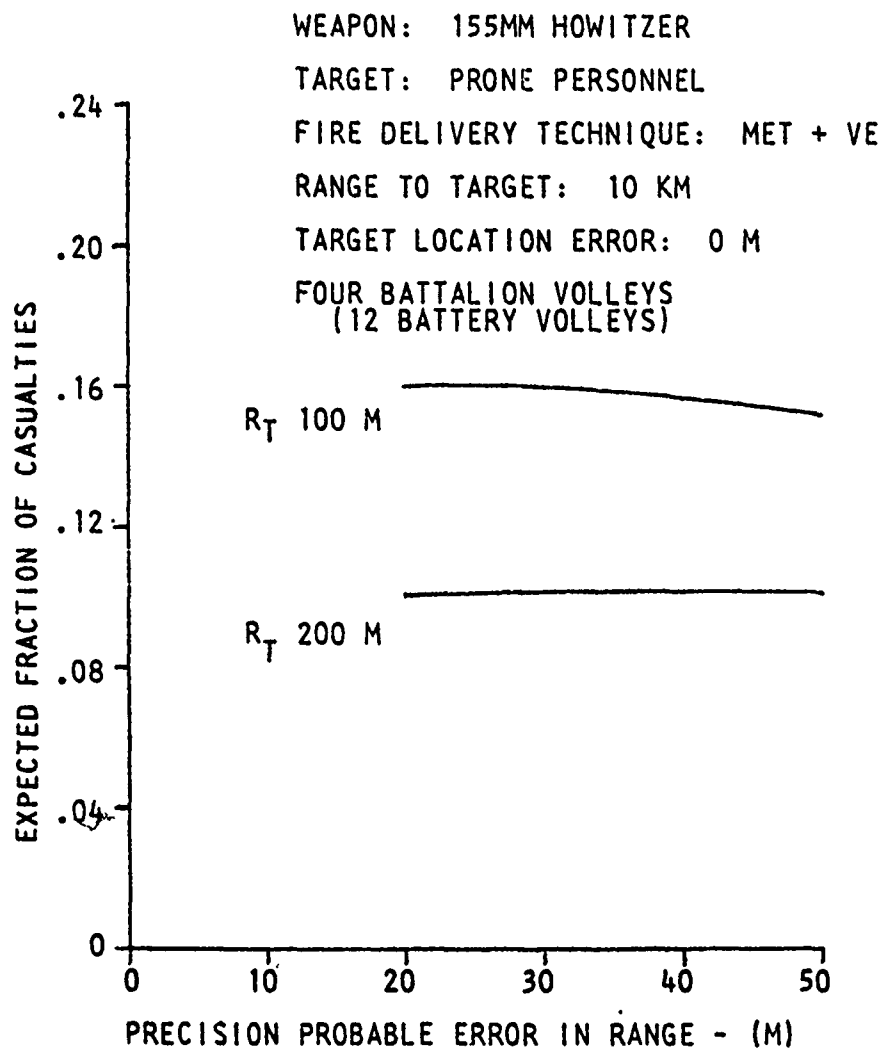


FIGURE 11

EFFECT OF PRECISION ERROR

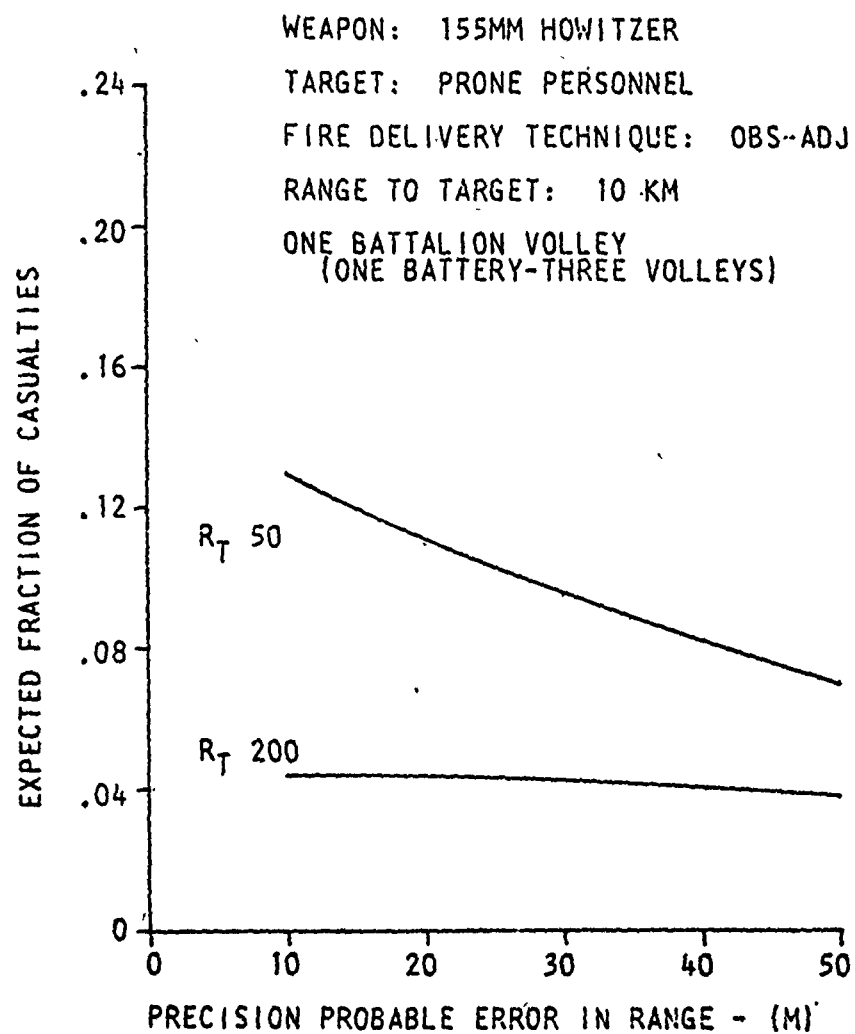


FIGURE 12

— UNSHIELDED
 - - - SHIELDED, SMOOTH FIELDS
 - - - SHIELDED, ROUGH FIELDS

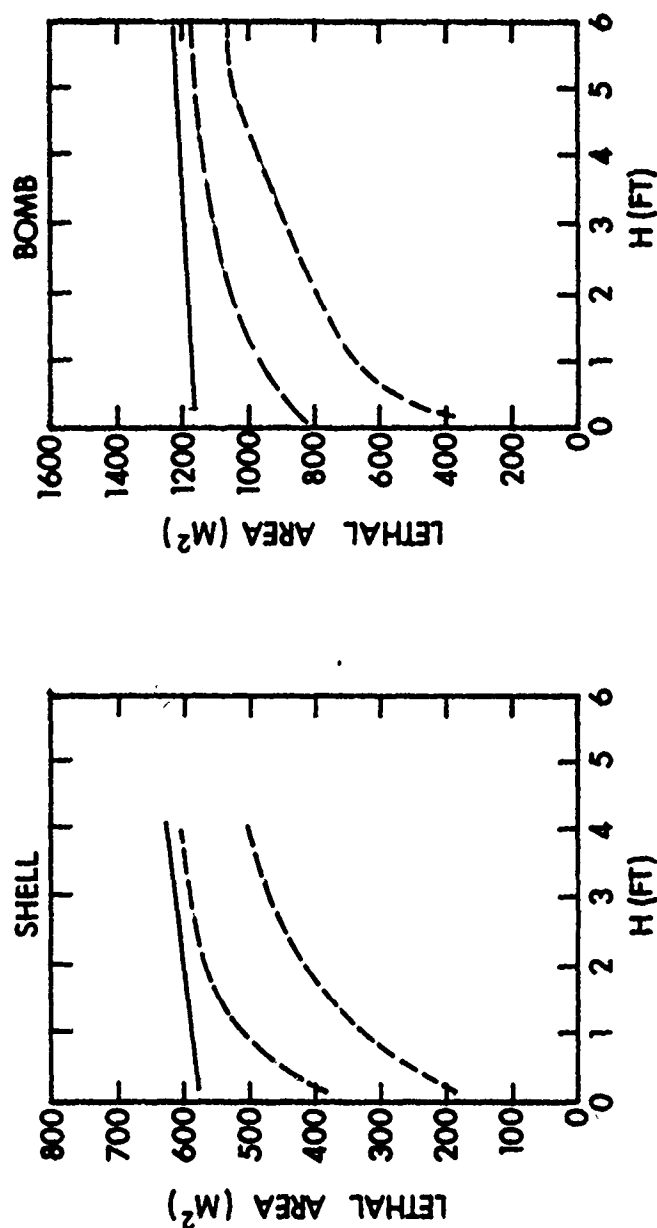


FIGURE 13

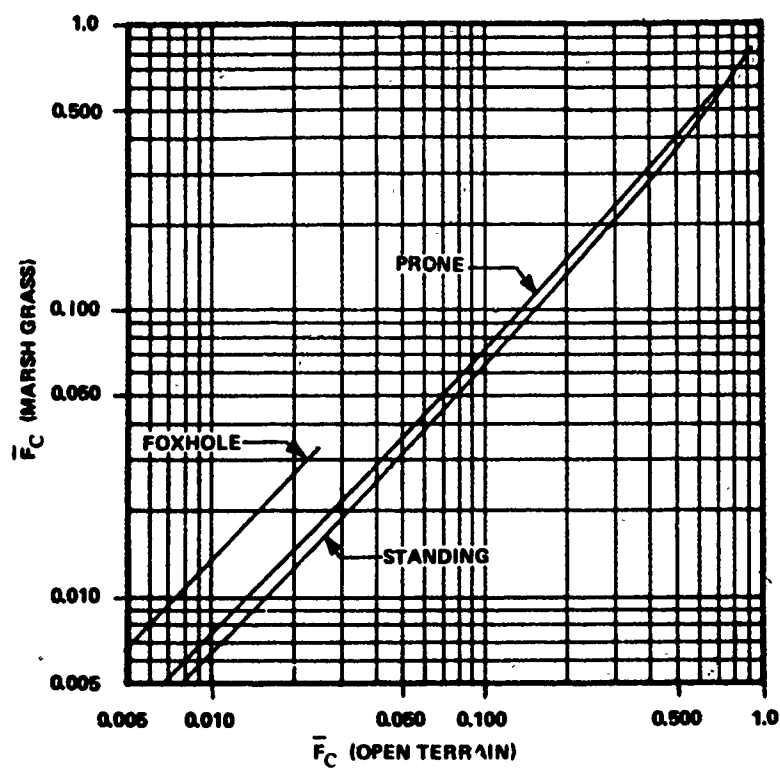


FIGURE 14

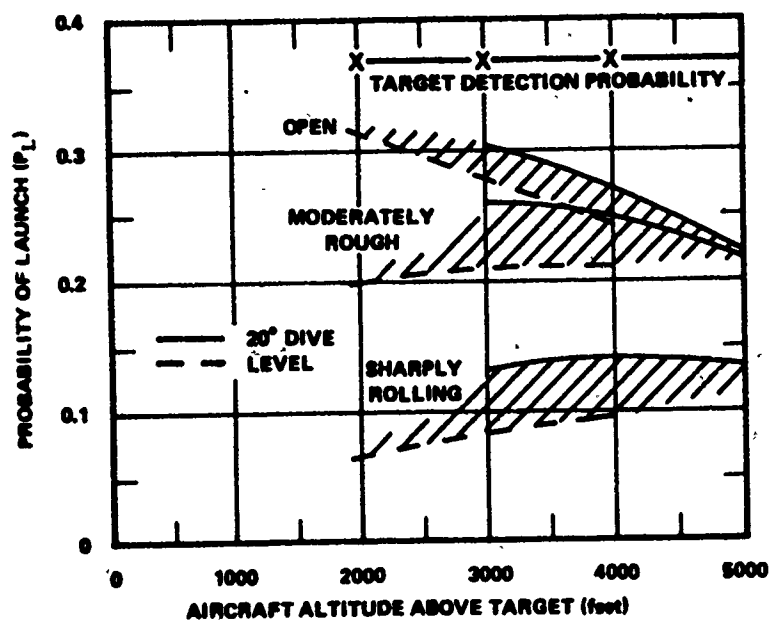


FIGURE 15

ARTILLERY AIMING TECHNIQUES STUDY

ILLUSTRATION OF AIMING TECHNIQUES

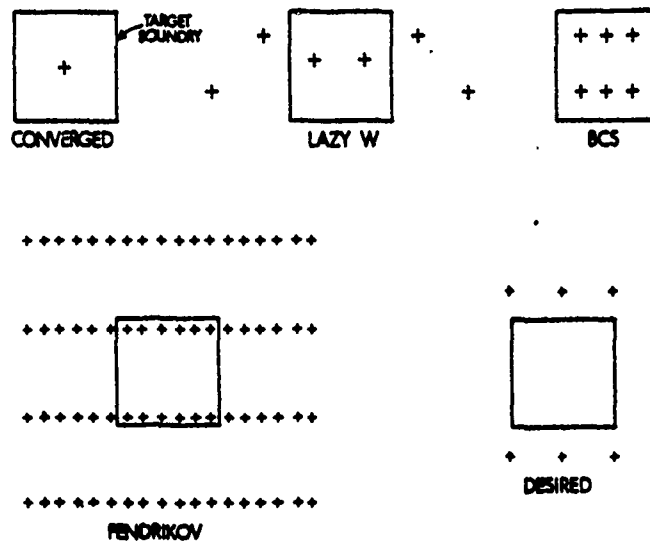


FIGURE 16

ALGORITHM FOR GENERATING DESIRED SPACING FOR SELECTION OF AIMING POINTS

TO DETERMINE RANGE SPACING - R^*

$$\Delta R = \sqrt{\left(\frac{1}{c^2-1}\right) \times \left(18 F^* E_{MPI_R}^2 + F^* L_R^2 - 45 E_{PREC_R}^2\right)}$$

TO DETERMINE SPACING BETWEEN FLANKING GUNS - D^*

$$\Delta D = \sqrt{\left(\frac{1}{d^2-1}\right) \times \left(24 F^* E_{MPI_d}^2 + F^* L_d^2 - 27 E_{PREC_d}^2\right)}$$

WHERE $F^* =$ DESIRED FRACTIONAL CASUALTY LEVEL (?) + .1
(i.e. $F^* = \bar{F} + .1$)

$E_{MPI_R}, E_{MPI_d} =$ MEAN POINT OF IMPACT ERROR IN RANGE OR DEFLECTION (PE)

$L_R, L_d =$ TARGET DIMENSION IN RANGE AND DEFLECTION

$E_{PREC_R}, E_{PREC_d} =$ PRECISION ERROR IN RANGE OR DEFLECTION (PE)

If value under square root sign is negative $\Delta^ = 0$

COMPARATIVE RESULTS OF ALTERNATIVE AIMING TECHNIQUES

TARGET SIZE - 100M x 100M
LOCATION ERROR - 150M CEP

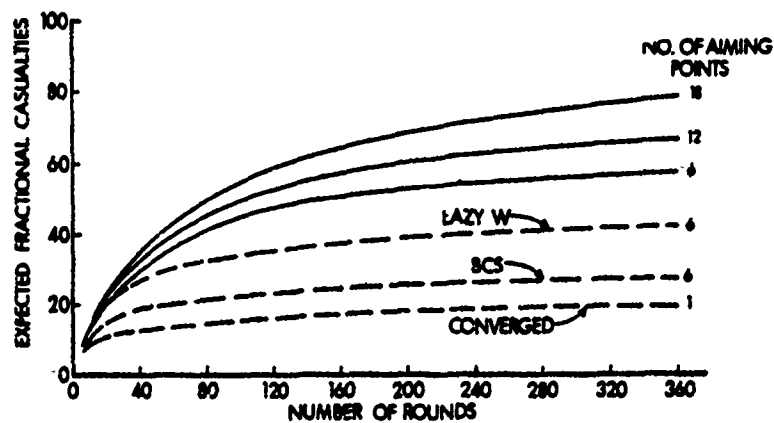
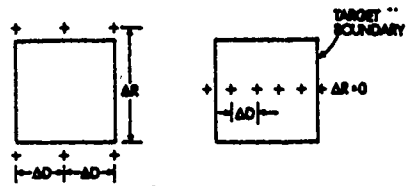


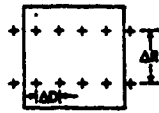
FIGURE 17

ILLUSTRATION OF PREFERRED AIMING PATTERNS:

ONE BATTERY



TWO BATTERIES



BATTALION

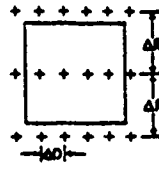


FIGURE 18

COMPARATIVE EFFECTIVENESS NO. OF ROUNDS FOR XX CASUALTIES PRONE PERSONNEL

TARGET SIZE (H)	HE AIMING TECHNIQUES			
	TLE (CEP H)	LAZY W	MOD FEND	PREFERRED
50 X 50	25	96	66	66
	75	114	96	96
	125	204	174	168
100 X 100	25	96	72	72
	75	126	96	96
	125	216	180	174
200 X 200	25	108	90	90
	75	156	126	126
	125	264	210	204
300 X 300	25	168	144	144
	75	216	180	180
	125	---	(284)	(258)
TOTALS		1764	1434	1416
PCT SAVINGS		---	19%	20%

FIGURE 19

**TRAINING OF PERSONNEL IN THE NIGERIAN
ARMY SIGNAL TRAINING SCHOOL**

TAIWO T. ABODUNDE and OLUWAFEMI FAYOMI

**Department of Computer Sciences
University of Lagos
Lagos, Nigeria**

ABSTRACT. The Nigerian Army believes that the most rewarding way of achieving efficiency in the system is to lay emphasis on good and regular training so that the acquisition of modern equipments could be put to maximum use. In pursuance of this belief, the Nigerian Army established a number of training schools few years ago to train its personnel. Among these training schools is the Nigerian Army Signal Training School (NASTS). One of the objectives of this school is the provision of basic and progressive military education and related practical training for the personnel of the corps of signals to enable it provide efficient and reliable communication system to the entire Nigerian Army.

This paper discusses the program of selecting and training signal corps in NASTS. The aim is to provide a data base which has hitherto been absent for future reference and planning. An attempt is made to provide future training schedules by the use of linear programming technique.

1. INTRODUCTION

This paper describes the training of personnel in the Nigerian Army Signal Training School (NASTS). The Nigerian Army in most respect, is just like any other organization in the developing countries today. A major problem in most of these organizations is lack of adequate planning in their developments. There is also lack of sufficient and well trained personnel which frequently results in poor strategic decisions. There is ample evidence that quality has been replaced with quantity in the infrastructural developments involving huge investments and it appears in most cases, as if there is no standard of measurement in an attempt to change the situation. Many developing countries have passed through some stages of growth without developments. This situation is common in every sector of the economy - The problems are made worse because of the significant pressure being put on the state of economy and rate of growth by the increasing technical complexity coupled with the increase in specialisation of personnel and equipment in developed countries. Expansion at a faster rate to meet the numerous challenges is also impeded by limited resources. But one aspect that needs to be kept going on a continuous basis is the training of personnel.

The Nigerian Army encourages scholarship and education within its rank and file. It believes that the most rewarding way of achieving efficiency in the system is to lay emphasis on proper and adequate training so that acquisition of modern equipments could be put to maximum use. In pursuance of this, the Army established a number of training schools few years ago to train its personnel. Among these training schools is the Nigerian Army Signal Training School (NASTS). One of the objectives of this school is the provision of basic and related practical training for the personnel of the corps of signals to enable it provide efficient and reliable communication system to the entire Nigerian Army.

One of the objectives of this study was to initiate, if necessary, a totally new training schedule that is expected to produce signal officers with more relevant training, provide greater assurance of experienced and good quality personnel, and reduce costs chargeable to training by a substantial amount. It was also our aim to examine measures of effectiveness, assess total training costs and compare

costs of alternative ways of training. Unfortunately, little reliable data was found to carry out a thorough analysis with respect to the above set out objectives.

The purpose of the above objectives was also to introduce, through this study, the methodology of Operational Research/Management Science into the Nigerian Army. For proper analysis of the problems facing a developing Army, there is a need for real training in economic, statistical, and computational tools to help increase the efficiency in planning and control in the system. At present, there is lack of sufficiently trained analytical or modelling personnel with respect to rapidly expanding requirements. At present, also, Operational Research/Management Science has not penetrated the planning process of the Nigerian Army and indeed of the planning process of Nigeria in any depth. But it is hoped that in the near future, it will make a significant contribution.

2. THE SIGNAL TRAINING SCHOOL

The NASTS is headed by a Commandant, who is usually a Major. The staff in the training school consists of both military and civilian instructors, as well as technicians. The purpose of establishing the school is to train the personnel of the corps of signals. After successful completion of the training period, the personnel are assigned to various corps of the Nigerian Army to provide efficient and reliable communication system to the entire Nigerian Army. Towards this end, the NASTS operates under the following guidelines:

- (i) To provide basic and progressive military education and related practical training for the personnel of the corps of signals;
- (ii) To conduct courses and test tradesmen in all signalling and allied trades;
- (iii) To provide signal courses to other corps in the Nigerian Army;
- (iv) To provide correspondence courses for soldiers who are one way or the other unable to attend regular courses in the training school;
- (v) To provide adequate facilities for interested personnel who may wish to sit for civil

professional examinations, and

- (vi) To advise the authorities on the introduction of other new signalling courses that may be of benefit to the corps and the Army in general.

The functions of the signal personnel are many and these depend on different specialisations.

3. THE PRESENT TRAINING SYSTEM

3.1 Basic Courses

The NASTS conducts courses in the following trades: Preliminary Courses; Basic Technical Courses; Radio Operator; Communication Centre Operator; Linesman; Crypto Rider; Despatch Rider; Driver Signallers; Draughtman Signallers; Electrical Technicians; Radio Technicians; Terminal Technicians; System Technicians; Instrument Technicians; Technical Storeman; Projectionists and Regimental Signallers. The Training in NASTS is conducted at three levels - The Preliminary Course (PC), The Basic Course (BC), and The Up-Grading Course.

3.1.1 The Preliminary Course

This Course is designed for soldiers who do not have adequate educational background. Basic English, Mathematics, as well as general knowledge required for the next stage of the training are taught to the students. The duration of this course is about 36 weeks.

3.1.2 The Basic Training

Successful students from the Preliminary Course are allowed to proceed to the Basic Training Course. This course is designed for students with general education in technical and operating trades. The duration of the course is about 27 weeks. Subjects offered include Electronics, Electricity and Magnetism, as well as Mathematics. It is after successful completion of this course that students are assigned to train in specialised trades like Radio, System, Instrument or Terminal Technician Course which usually last for about 32 weeks.

After this training, successful technicians are posted to regiments or brigades to acquire practical experience.

3.1.3 The Up-Grading Course

The Up-Grading Course is designed for the students who have passed out of the NASTS and have been posted to the field to have practical experience at some unit level. After a specific period of field experience, they are recalled back to the NASTS to take an Up-Grading Test. It has been observed that the rate of reporting back to the school for the upgrading examinations is not encouraging.

3.1.4 The Pre-University and University Courses

Apart from the above courses, there is also the Pre-University Course for officers from the Nigerian Defence Academy (NDA). These officers are those who successfully completed their military training and are interested in sciences, especially Physics and Mathematics. The objective of the course is to prepare the officers for a specially arranged Diploma Course in Telecommunication Engineering with the Faculty of Technology of the University of Ife, Nigeria. The Pre-University Course is usually from January to September of each year. Successful officers are then allowed to proceed to the University for the two-year Diploma Course. The subjects offered in the University include Thermodynamics, Electronic Engineering, Computer Science, Engineering Drawing, System Control, Industrial Engineering, Industrial Management as well as a long period of Industrial Training.

The NASTS has recorded some degree of success over the last few years despite inadequate staffing, insufficient classroom facilities, lack of adequate training equipments, as well as frequent change of command.

4. PROJECTED TRAINING LOAD

Little reliable data was found to estimate the number of students that would require training in the next few years. The wide fluctuation of this number over the past few years showed some inconsistency in the recruitment procedures. It was our intension to use the most elementary method of forecasting, i.e. linear smoothing and projection, to estimate the future training load by using a difference equation relating needed recruits to the major relevant variables of numbers of trained signal personnel required, and training wastage. But data on the last two variables were not available.

The results of thirty courses undertaken in the NASTS from 1975 to 1977 as well as the number of various grades of technicians that were turned out were compiled. A rough estimate of the expected intake for the following year, 1978, was obtained from the trend of the past three years. The above data as well as unavailable information needed for future analysis were compiled. A computer program for updating this record every year was written and run. It is hoped that this would be of help to the NASTS in its future plans and also for future studies.

5. THE SCHEDULING PROBLEM

The second task was to identify and evaluate the alternative methods available to satisfy present and future training needs. The key issues are the cost and duration of training, the ability of the training system to handle various loads. Our effort was restricted to the scheduling problem called the "Smoothing Problem". For more details, refer to {2}, {3}.

In it, we let R_1, R_2, \dots, R_n be the number of graduates desired in the different categories of trades in the NASTS. If these targets are anticipated, some of the graduates will be temporarily in excess and some expenses will be incurred in providing for them till they are needed. It is expensive to vary the size of successive classes to meet requirements exactly on schedule. In this case, the cost of increasing the size of a class is measured as proportional to the amount of increase from the preceding class. The cost of a decrease can be neglected.

To set up a mathematical model for the determination of an efficient compromise between these conflicting procedures, we let X_j be the number of graduates from j th year, S_j be the number of excess graduates on hand after fulfilling the requirements in that year. Then

$$X_1 + X_2 + \dots + X_j = R_1 + R_2 + \dots + R_j + S_j$$

$$(j=1, 2, \dots, n) \quad (1)$$

The cost of providing for the excesses is proportional to the number of excess graduates. The cost of necessary increase in the rate of training is proportional to

$$(X_2 - X_1) + \dots + (X_n - X_{n-1})$$

where $X_2 - X_1$ is the increase from the first to the second class, etc.

The objective function can therefore be written as

$$S_1 + S_2 + \dots + S_n + K\{(X_2 - X_1) + \dots + (X_n - X_{n-1})\}$$

$$(2)$$

This should be a minimum. That is, the objective of the exercise is to minimize the total cost of training and maintaining excess graduates from a particular training course. K in the equation measures the cost of one unit increase in class size, relative to the cost of carrying one excess of graduate for one period.

The above system because of the form of the objective function, is not quite a linear programming problem. It is converted to the strict linear programming form by a device due to Dantzig [3]. We can thus write

$$X_2 - X_1 = Y_2 - Z_2$$

$$X_3 - X_2 = Y_3 - Z_3$$

$$\vdots$$

$$X_n - X_{n-1} = Y_n - Z_n$$

where the Y_j and Z_j are non-negative.

These equations are added to the original constraints and they replace the original objective function by

$$S_1 + S_2 + \dots + S_n + K(Y_2 + \dots + Y_n) \quad (3)$$

For an optimum solution, either Y_j or Z_j will be zero for every j , so that linear programming model will give the same answers as the other even though the two are not strictly equivalent. The complete model is then

Minimize $\sum S_j + K \sum Y_j$

subject to $S_1 + S_2 + \dots + S_j = R_1 + R_2 + \dots + R_j + S_j$
($j=1,2,\dots,n$)

$X_j - X_{j-1} = Y_j - Z_j$ ($j=2,3,\dots,n$)

The quantity k in the objective function is referred to as the relative cost of fluctuation versus excess or the relative disutility.

In applying the above, the three years of graduate production records from NASTS were used. Because of the inconsistencies in the data available, some of the targets were arbitrary, thereby making the whole exercise an academic one. Nevertheless, the authorities concerned were aware of the kind of information required as well as the objective of the exercise.

6. CONCLUSION

There is no doubt that little has been achieved with respect to how to set our objectives because of little reliable data available. Nevertheless, the authorities of the Nigerian Army Signal Training School are aware of the need to keep and supply reliable information for proper analysis that may improve the training system. Because of their encouraging response, it would be possible in the next few years to present the results of our study in details.

The problem encountered here is common in many organizations in the developing countries where it is difficult to establish what the true position is. It is the duty of an Operational Research team to encourage an improvement on this.

7. REFERENCES

- 1 FAYOMI, O.: SOME ASPECTS OF MILITARY APPLICATION
 OF LINEAR PROGRAMMING.
 - Department of Computer Sciences,
 University of Lagos,
 Lagos, Nigeria, 1978
- 2 DANTZIG, G.B.: LINEAR PROGRAMMING AND EXTENSION
 - Princeton University Press, 1963
- 3 JACOBS, W.: MILITARY APPLICATIONS OF LINEAR
 PROGRAMMING.
 - Proceedings of the Second Symposium in
 Linear Programming.

A DECISION-THEORETIC APPROACH TO
EVALUATING EFFECTIVENESS OF RECONNAISSANCE
SYSTEMS IN A TARGET ACQUISITION ROLE

JOEL A. HASSELL

The BDM Corporation
7915 Jones Branch Drive
McLean, VA 22102, U.S.A.

ABSTRACT. The use of reconnaissance assets to develop targeting data is a key factor in the capability to effectively deliver high value munitions on specific classes of targets. As a part of recent research efforts in this area, a methodology was developed which permits assessment of the use of reconnaissance systems in a targeting role and provides a means of evaluating the relative effectiveness of the various systems.

The methodology employs a decision-theoretic approach in a Monte Carlo simulation which permits consideration of both the interaction of the reconnaissance systems and the target environments, and the human factors involved in data interpretation and decision making. The approach takes advantage of what is known about the composition of individual target classes as well as the information available on target arrays in a given tactical situation. Information fusion from multiple reconnaissance systems and value judgments relative to the utility of various combinations of decisions and states of nature are included.

The methodology will not only permit evaluation of the capability of individual reconnaissance systems but will provide insights into operational issues. Such issues include questions regarding how best to employ existing reconnaissance systems, the value of additional information versus the time required to develop that information, the classes of targets which can best be exploited using reconnaissance systems, and the performance characteristics which should be designed into new systems.

1. INTRODUCTION

The effective use of limited numbers of high value, special purpose munitions depends upon the capability to deliver those munitions against specific classes of targets. Hence, the capability to locate and identify targets becomes a key factor in utilization of such weapons. The target acquisition process is defined as those activities involved in locating target elements and identifying the located elements as targets. A target may be defined as a single element such as a tank or as any configuration of elements such as a tank company. Identification may be as coarse as recognition of something which is not a part of the environment or as fine as recognition of a specific type of unit. However, the fundamental processes involved are the same regardless of the targets sought or the identification capability.

This paper describes the methodology developed to model the target acquisition process. The following major factors were considered:

- Geographic area;
- Target types;
- Sensor systems;
- Data processing;
- Data correlation;
- Decision making.

In order to facilitate development of the methodology it was assumed that:

- The target elements are vehicles (trucks, ADCs, tanks);
- The targets of interest are company-size units;
- The sensor systems are airborne systems;
- The data processing and correlation functions are performed by human operators.

However, the methodology has general applicability and is not limited by these assumptions. Additional assumptions will be discussed as they come into play in the development. The discussion will be divided into two sections, the detection of elements and the identification of targets.

2. DETECTION OF ELEMENTS

Suppose that in a given target region the possible targets are T_1, T_2, \dots, T_m , consisting of target elements E_1, E_2, \dots, E_n . The sequence of events involved in the detection of elements is shown in Figure 1. Depending on the sensor system in use, the sensor report may or may not include identification

of the elements E_i detected. It will, however, include the number of elements detected either in total (no recognition) or by element type (recognition).

For a specific target of type i with composition

$$T_i: N_{i1}E_1, N_{i2}E_2, \dots, N_{in}E_n, N_{ij} \geq 0, \quad (1)$$

a sensor report including identification of elements will take the form

$$R = [r_1E_1, r_2E_2, \dots, r_nE_n], r_j \geq 0, \quad (2)$$

where r_j is the number of elements of type j reported. Note that while the N_{ij} are double subscripted to indicate target type and element type, the r_j reflect only the element type since the target type is unknown to the sensor and the processor. For convenience of notation the E_i will be dropped and the sensor reports written

$$R = [r_1, r_2, \dots, r_n]. \quad (3)$$

In the trivial case with no recognition of element types, the sensor report is a single number of detections, r .

Figure 2 illustrates the detection process for a single target element E_i . The notation used is as follows:

- PUN - probability that the element is unmasked;
- PSR - probability of sensor return from the element;
- PD - probability that the sensor return is detected;
- PIDC - probability that the detected return is correctly identified as element type E_i ;
- PIDI - probability that the detected return is incorrectly identified as element type E_j , $j=1, \dots, n$; $j \neq i$.

Nodes 1, 2 and 3 represent cases in which the specific E_i will not be included in the sensor report because it is masked, not sensed, or undetected. Node 4 is the final node for a sensor which does not permit recognition of element types; the detected element will be included in the sensor report. Node 5 represents identification of the element incorrectly as a false alarm; it is not reported. Nodes 6 and 7 represent

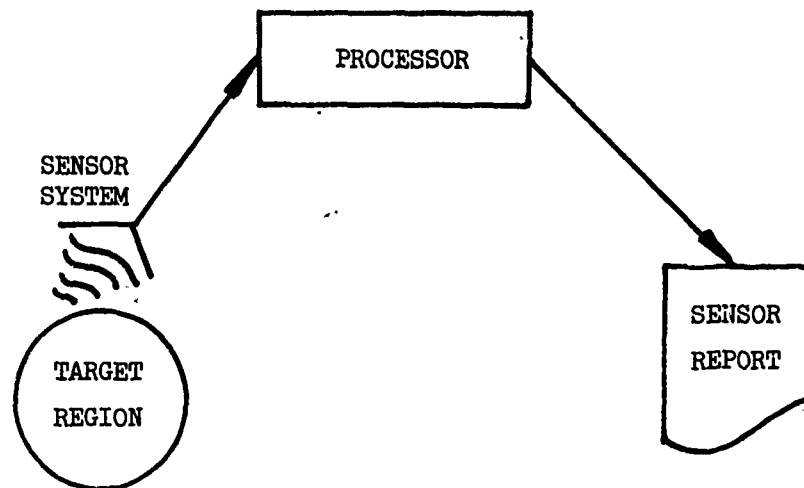


Figure 1. Detection Process.

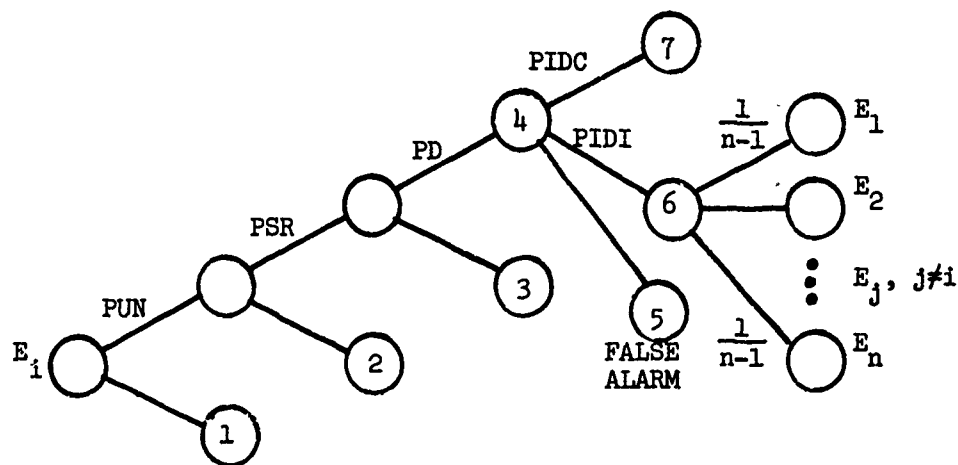


Figure 2. Detection of a Target Element.

inclusion of E_i in the sensor report identified either incorrectly or correctly. The probability $1/(n-1)$ assumes uniformity in identifying the element type incorrectly; however, any other distribution could be used.

An additional factor in the composition of the sensor reports is the occurrence of false alarms. A false alarm is a return from something other than a target element. Two assumptions concerning false alarms have been made:

- The distribution of false alarms is Poisson with the mean rate of occurrence, λ , a function of the geographic area and the sensor system;
- Some false alarms will be recognized as false alarms, with probability PFI, and others reported as target elements, with probability $1-PFI$.

Figure 3 depicts the false alarm process. Node 1 reflects the false alarm being recognized as a false alarm while the branches off Node 2 represent the false alarm being reported as an element E_i . The probability $1/n$ assumes uniformity in identifying false alarms as element types; again, any other distribution could be used.

In a sensor report

$$R = [r_1, r_2, \dots, r_n],$$

the r_i may include elements E_i correctly identified, elements E_j incorrectly identified as type E_i , and false alarms reported as elements E_i . A report from a sensor system incapable of recognizing elements by type is simply the total of all detected returns and reported false alarms.

3. IDENTIFICATION OF TARGETS

Given a sensor report R of the form

$$R = [r_1, r_2, \dots, r_n],$$

what can be said about the target from which it came? Simplistically one might look at the report and decide that since there are apparently r_1 elements of type 1, r_2 elements of type 2, and so on, which have been detected it is probably a type T_k target. However, there is additional information which can improve the probability of correct identification.

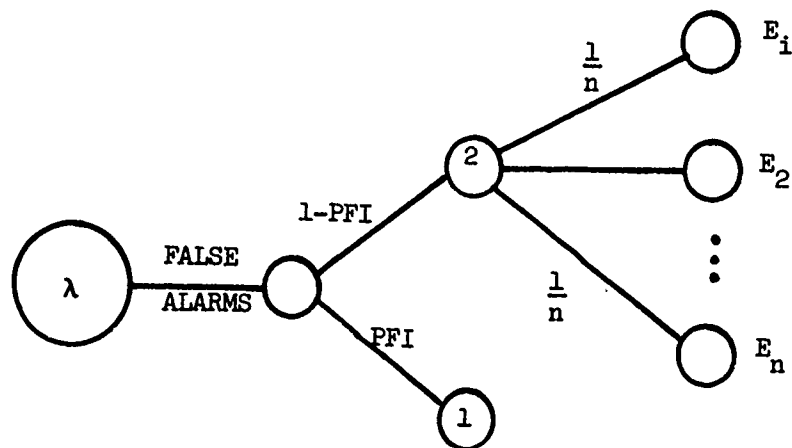


Figure 3. False Alarm Process.

Suppose that the target region may contain targets of type T_1, T_2, \dots, T_m . It is assumed that a given sensor report R is generated from false alarms and/or elements detected which belong to at most one target,

$$T_i: N_{i1}, N_{i2}, \dots, N_{in} \quad (4)$$

The report R can be decomposed as follows. One component is the vector R_0 of reported false alarms:

$$R_0 = [r_{10}, r_{20}, \dots, r_{n0}] \quad (5)$$

If elements from a target of type T_i were also detected, then for each element type E_j there is a vector,

$$R_j = [r_{1j}, r_{2j}, \dots, r_{nj}] \quad (6)$$

which results from sensor responses to the N_{ij} elements of type E_j in the target. The report R is the sum of all these vectors:

$$R = \sum_{j=0}^n R_j, \quad r_k = \sum_{j=0}^n r_{kj} \quad (7)$$

The vector R_j is generated from N_{ij} trials on elements of type j . It is a random vector with a multinomial distribution:

$$M(R_j; N_{ij}, \langle p_0, p_1, \dots, p_n \rangle) =$$

$$\frac{N_{ij}!}{(N_{ij} - \sum_{k=1}^n r_{kj})! \prod_{k=1}^n (r_{kj}!)} \cdot p_0^{(N_{ij} - \sum_{k=1}^n r_{kj})} \cdot \prod_{k=1}^n p_k^{r_{kj}} \quad (8)$$

where

$$\sum_{k=1}^n r_{kj} \leq N_{ij}$$

$$p_j = PUN \cdot PSR \cdot PD \cdot PIDC$$

$$p_k = PUN \cdot PSR \cdot PD \cdot (PIDI/(n-1)), k \neq j$$

$$p_0 = 1 - \sum_{k=1}^n p_k.$$

The distribution of R_0 can be derived as follows. First, the probability of N false alarms is

$$P(N; \lambda) = e^{-\lambda} \frac{\lambda^N}{N!}. \quad (9)$$

Let

Θ_k = probability of reporting a false alarm as an element of type k , $k=1, \dots, n$,
 Θ_0 = PFI = probability of recognizing a false alarm as false.

Given N false alarms, the probability of reporting

$$R_0 = [r_{10}, r_{20}, \dots, r_{n0}] \quad (10)$$

and recognizing r_{00} sensor responses as false, where

$$N = \sum_{k=0}^n r_{ko},$$

is given by the multinomial distribution

$$P[R_0|N] = N! \prod_{k=0}^n \frac{\Theta_k^{r_{ko}}}{r_{ko}!}. \quad (11)$$

Therefore the unconditional probability of R_0 together with r_{00} false alarms recognized as false is

$$P[(r_{00})UR_0] = \prod_{k=0}^n \left(e^{-\lambda \Theta_k} (\lambda \Theta_k)^{r_{ko}} / r_{ko}! \right). \quad (12)$$

The probability of R_0 is then obtained by the infinite summation

$$\begin{aligned}
 P[R_0] &= \sum_{r_{00}=0}^{\infty} P[r_{00}, R_0] \\
 &= \prod_{k=1}^n \left(e^{-\lambda\theta_k} (\lambda\theta_k)^{r_{ko}} / r_{ko}! \right). \quad (13)
 \end{aligned}$$

Thus the components of the vector R_0 have independent Poisson distributions with means $\lambda\theta_k$, $k=1, \dots, n$.

It is natural to assume that the processes generating the random vectors R_0, R_1, \dots, R_n are independent. This implies that the probability distribution of R , given the fact that a target of type i was observed, is the convolution of the distributions of $R_0, R_1, R_2, \dots, R_n$ derived above. In the case where the sensor report was generated solely from false alarms, the distribution of R coincides with that of R_0 .

Suppose now that a sensor report

$$R = [r_1, r_2, \dots, r_n]$$

is received. The preceding calculations enable one to compute the conditional probabilities $P[R|T_i]$, the probability of the report coming from a target of type i , $i=1, \dots, m$, and $P[R|F]$, the probability of the report being generated solely from false alarms. If $P[R|F]$ is the largest of these $m+1$ conditional probabilities, one might conclude that no target was observed. Likewise, if $P[R|T_h]$ is largest, then one might decide to report a target of type h . However, such a maximum likelihood decision rule fails to take into account additional information which can be used to improve the accuracy of the target reports. There may have been previous target reports from the same location, there may be an overall intelligence estimate concerning the region, and the decision maker may have information derived from his experience and judgment which bears on the problem.

Rather than the probabilities $P[R|T_i]$ and $P[R|F]$, the probabilities $P[T_i|R]$ and $P[F|R]$ would be of greater value in the

target identification process. That is, given the sensor report R has occurred, what are the probabilities that it came from target type i or from false alarms? From Bayes theorem:

$$P[T_i | R] = \frac{P[R|T_i]P[T_i]}{P[R|F]P[F] + \sum_{a=1}^m P[R|T_a]P[T_a]} \quad (14)$$

$$P[F | R] = \frac{P[R|F]P[F]}{P[R|F]P[F] + \sum_{a=1}^m P[R|T_a]P[T_a]} \quad (15)$$

Since $P[R|F]$ and $P[R|T_i]$, $i=1, \dots, m$, have been computed, only the probabilities $P[F]$ and $P[T_i]$ are required. These are called prior probabilities.

As an illustration of how "priors" might be obtained, consider the following situation. A sensor mission is to be flown over a specific region of interest to the decision maker. There are no previous target reports which are relevant to the particular time and place. The decision maker is concerned with locating company-sized targets and there is an intelligence estimate which suggests that there is an armor regiment somewhere within the region. It is known that an armor regiment generally has nine tank companies, one reconnaissance company, an ADA battery, three battalion headquarters, a regimental headquarters, and various service units. Based on his experience and using map analysis, the decision maker estimates the number of each unit type he would expect to find in that area and their possible locations. Using this data and information on unit size in terms of area occupied, it is possible to compute a probability of finding a particular target type T_i in any given location. This is the prior probability $P[T_i]$. One can also compute a probability of finding no targets in a given location. This serves as the prior $P[F]$. The composition of each target type in terms of elements such as tanks, APCs, trucks, and so on is also generally known. This determines the quantities N_{ij} used in computing $P[R|T_i]$.

Having established prior probabilities, it is now possible to compute the conditional probabilities $P[T_i|R]$ and $P[F|R]$.

If $P[F|R]$ is largest among these conditional probabilities, the decision maker might conclude that no target was acquired. On the other hand, if $P[T_h|R]$ is largest, a target of type h might be reported. However, the decision maker may wish to associate utility values with reports of specific target types to reflect targeting priorities. For a target of type i the expected utility value is

$$EV(T_i) = \sum_{j=1}^m P[T_j|R] U_{ij} \quad (16)$$

where U_{ij} is the utility associated with reporting target type i when the report resulted from target type j .

Further, since the decision maker may expend limited resources based on the target reports it may be desirable to establish a minimum expected value threshold, $EVTHR$, such that a target of type i is reported only where

$$EV(T_i) > EV(T_j) \quad j=1, \dots, m \quad i \neq j \quad (17)$$

and

$$EV(T_i) \geq EVTHR.$$

The preceding discussion has been based upon a sensor report from a single sensor system. The methodology was also generalized to treat the situation in which sensor reports from several different sensor systems are available.

Suppose that s different sensor systems produce the sensor reports $R^{(1)}, R^{(2)}, \dots, R^{(s)}$ from a given location which may or may not contain a target. The computations outlined above enable one to calculate the conditional probabilities $P[R^{(a)}|F]$ and $P[R^{(a)}|T_i]$ for $a=1, \dots, s$ and $i=1, \dots, m$. It should be noted that various parameters such as λ , PSR , $PIDI$, etc., depend on the sensor system being used. It is assumed that the sensor systems operate independently. If R denotes the composite of all the sensor reports, then

$$P[R|F] = \prod_{a=1}^S P[R^{(a)}|F], \quad (18)$$

$$P[R|T_i] = \prod_{a=1}^S P[R^{(a)}|T_i]. \quad (19)$$

Using these conditional probabilities the $P[F|R]$, and $P[T_i|R]$ and $EV(T_i)$ are computed exactly as before.

4. CONCLUSION

The methodology discussed above has been implemented in a Monte Carlo simulation which permits modeling the acquisition process considering the use of multiple sensor systems against a user defined target environment. The simulation output includes the expected number of target reports for each target type, the expected number of those reports which are correct, the expected number of targets not reported for each target type and the 90 percent confidence limits for correct reports.

Work is in progress to enhance the simulation. Two specific areas of interest are computing revised prior probabilities as a function of the sequential target reports and developing the techniques to treat the dynamics of mobile targets.

MULTIDIMENSIONAL PARAMETRIC ANALYSIS USING RESPONSE SURFACE
METHODOLOGY AND MATHEMATICAL PROGRAMMING AS APPLIED TO
MILITARY PROBLEMS

PALMER W. SMITH
HQ JUSMAG-K (MKDA-D)
APO San Francisco 96302

JOSEPH M. MELLICHAMP
MANAGEMENT SCIENCE AND OPERATIONS MANAGEMENT DEPT
P.O. Box J
University of Alabama
University, Alabama 35486

ABSTRACT. Uncertainty exists in military force structure and related analyses because of a lack of knowledge of the "true" values of pertinent parameters and an inability to explicitly define relationships between all parameters. In order to gain some insight into the uncertainty, the classical approach is to conduct a sensitivity analysis by varying the value of a selected parameter. Examining a model's solution over a wide range of a parameter's values, provides a one-dimensional picture to the decision maker of the sensitivity of the solution to parameter variation.

The single parameter sensitivity technique, however, has major disadvantages. This paper presents the recently developed multidimensional parametric analysis methodology which provides an expanded capability for conducting a more valuable analysis in a complex environment. It provides a picture of what is happening within the model being used for the study. It also provides insight into the relationship among the factors under study. "What if" analyses can be conducted economically and in real time without the necessity to obtain new computer outputs. Derived contributions of each parameter to the value of the measure of effectiveness and ratio comparisons of the different coefficients of the model parameters provide a new measure of effectiveness for comparing the worth and capability of one system vs another within the context of the model.

This paper presents and demonstrates the use of the methodology as applied to a simulated military force structure analysis problem. However, the methodology can be used with any decision model, linear or non-linear in form.

1. INTRODUCTION

Uncertainty exists in military force structure analyses because of a lack of knowledge of the "true" values of pertinent parameters used in the analysts's models and an inability to explicitly define relationships between all parameters. In order to gain some insight into the uncertainty, the classical approach is to conduct a sensitivity analysis by varying the assumed value of a selected parameter. Examining a model's solution over a wide range of a parameter's values, provides a one-dimensional picture to the decision maker of the sensitivity of the solution to parameter variation. This technique is well established in military operations research and systems analysis and has been valuable in illuminating the criticality of parameter values and certain assumptions about parameter relationships and assumptions about weapon systems characteristics and the environment in which they are used.

The single parameter sensitivity technique, however, has three major disadvantages. First, only one parameter can be varied effectively at a time. Secondly, this method provides no numerical measure or ranking of the importance of a parameter to the end solution. And, thirdly, no information is provided about the interrelationships between the important factors under study.

Figure 1 shows an example of the standard one-dimensional sensitivity analysis.¹

In this example the relationship between the number of B2s and this parameter's effect on Damage Expectancy (DE) is under investigation. Varying the number of B2s over the range yields the graph in Figure 1. When the relationship is linear as depicted by the straight line a very simple measure of the average contribution of a B2 to

1. Bombers, Intercontinental Ballistic Missiles (ICBM's) and Submarines are the categories of weapons systems used in this paper: Bomber-type 1 (B1), Bomber-type 2 (B2), ICBM-type 1 (M1), ICBM-type 2 (M2), and Submarine-type 1 (S1). Hypothetical numbers of these systems are used and allocated against a target data base to illustrate the usefulness of the methodology.

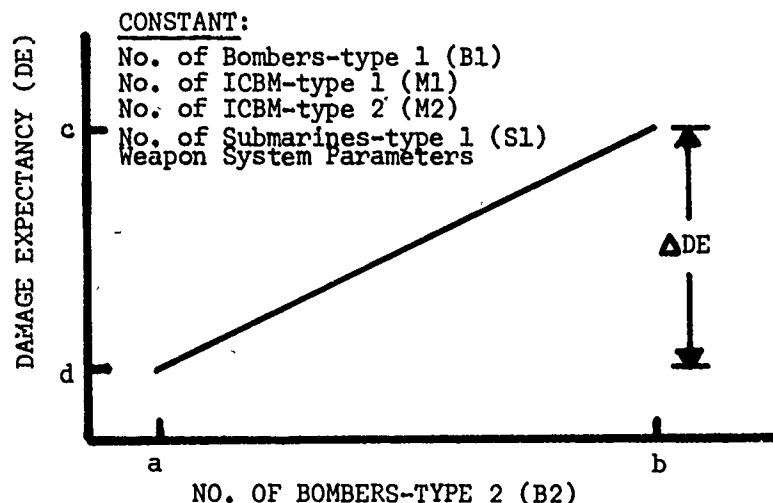


FIG. 1

Example of a One-Dimensional Sensitivity Analysis

total force DE is $(b-a)$ divided by the delta DE, $(c-d)$. This relationship is for a constant value of other system numbers and characteristics such as probability of penetration (PP), fixed weapon load, weapon yields, Circular Error Probable (CEP), weapon system reliability (WSR), probability of launch survivability (PLS), etc. The types of questions that cannot be answered reasonably with this one-dimensional sensitivity analysis include:

What is the average contribution of a B2 to total force DE over a range of its other systems characteristics?

How does the B2 contribution to total DE vary with respect to the values of B2 system characteristics that were held fixed?

What happens to the DE as several of the B2 weapon systems characteristics change at one time rather than one at a time in isolation?

What are the intra-parameter relationships, such as the interactions between the weapon system characteristics, within the model?

Likewise, if several different types of weapon systems are included in the analysis, there are questions which deal with the interactions of the different weapon systems which cannot be investigated fully with one-dimensional sensitivity analysis. Example of such questions are:

What are the relationships between parameters such as B2 probability of arrival (PA) over target, the desired total force DE, and the number of B1 bombers?

What is the potential impact of a significantly reduced submarine missile's CEP on the number of weapons required for each of the number of B2 bombers?

How does the PLS of an advanced ICBM affect the relationships obtained for the previous question?

Answers to these types of questions require a multidimensional sensitivity analysis, a simultaneous picture of the concurrent variations of DE to several parameters. Such a multidimensional picture is not practical using standard one-dimensional sensitivity analyses. However, a multidimensional analysis is possible utilizing the concept of response surface methodology [1] [3]. The response surface provides a picture of the relationship between the primary measure of interest and the input parameters (usually called factors) to the study.

2. THE RESPONSE SURFACE

Figure 2 is an example of a three-dimensional response surface. The shaded area is the response surface which consists of all of the values of the response, Y , that correspond to the possible combinations of the values of the factors, X_1 and X_2 , shown on the other axes. Assume that there exists a postulated relationship between the response variable and the factors as given in equation (1).

$$Y = b_0 + b_1 (X_1) + b_2 (X_2) + b_3 (X_1^2) + b_4 (X_2^2) \quad (1)$$

Assume also, that the value of X_1 ranges from 0 to 4 and that of X_2 from 0 to 4. Using equation (1), the value of Y can be calculated for each possible combination of X_1 and X_2 values (X_1, X_2). Plotting all the values of Y

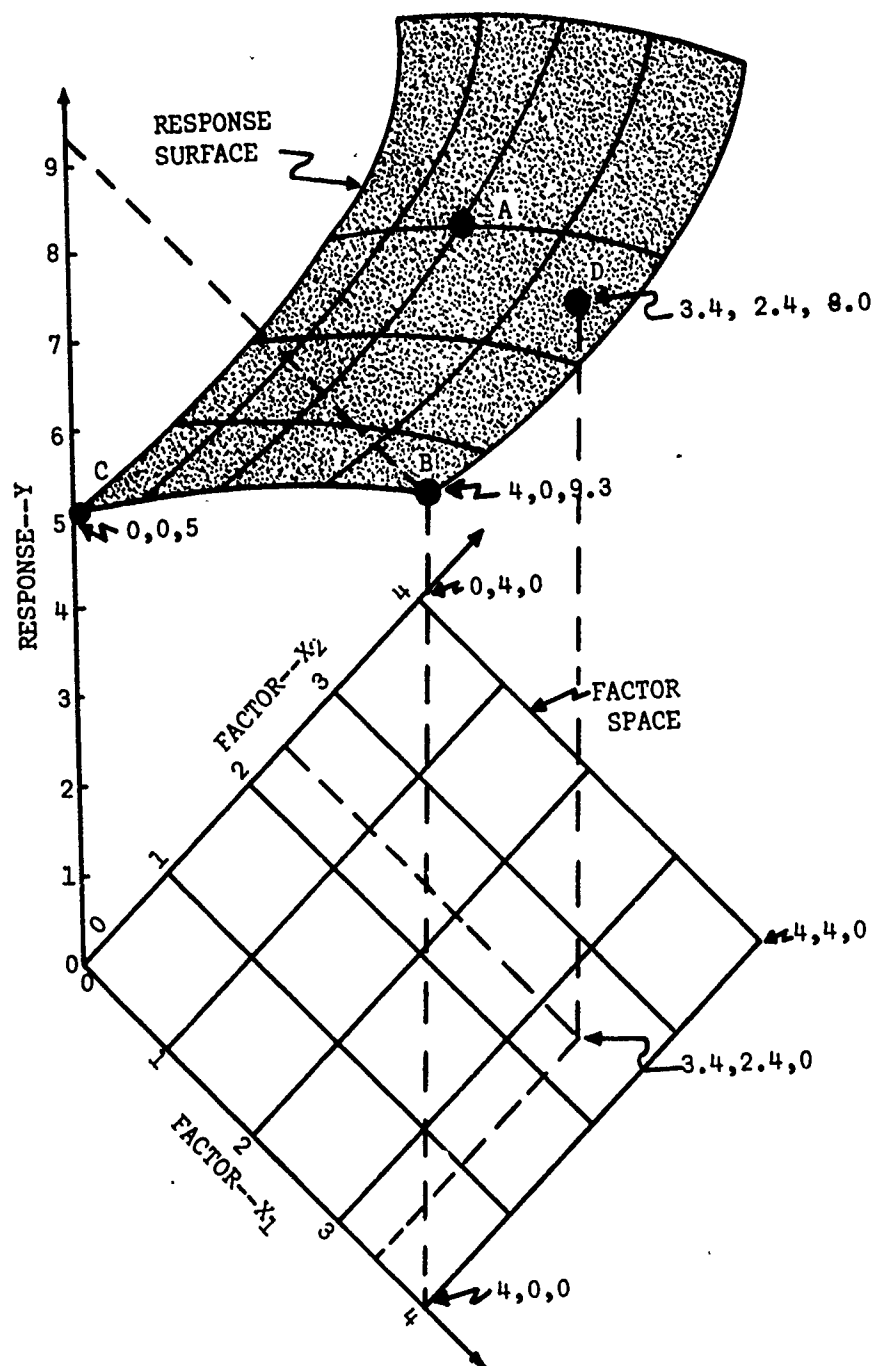


FIG. 2

A Response Surface

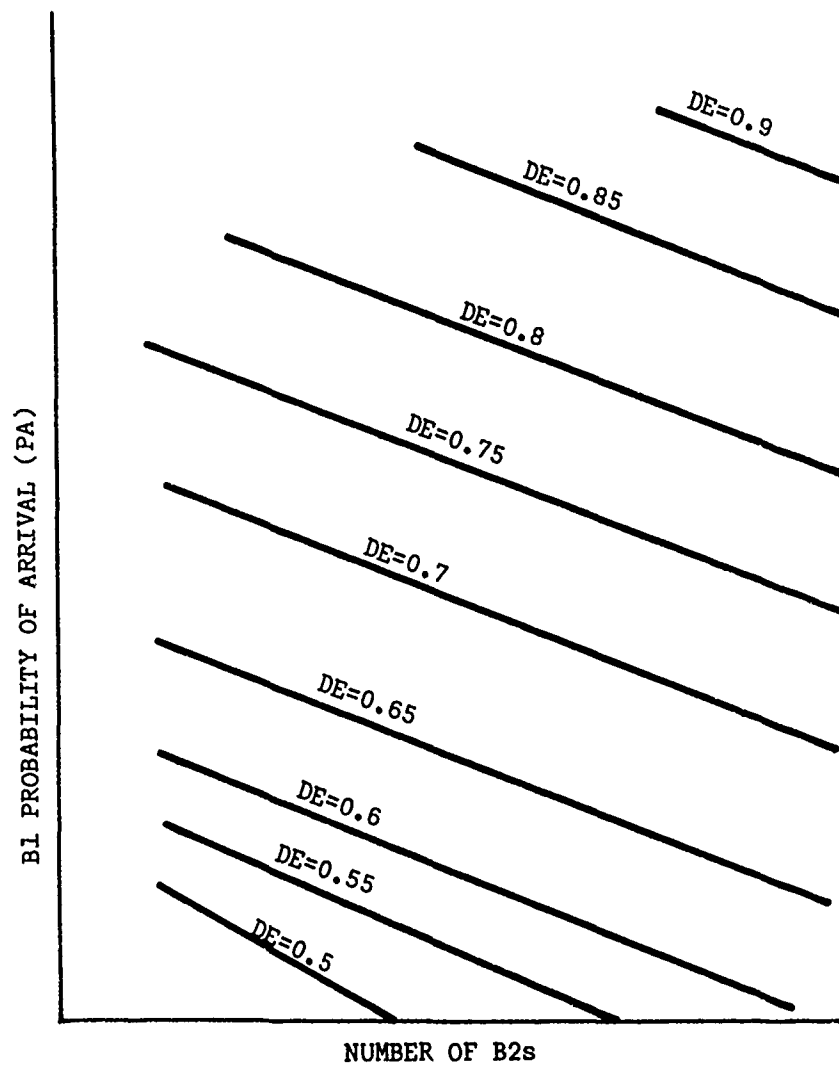
for the infinite number of possible combinations (X_1 , X_2) gives the response surface shown in Figure 2.

In almost all practical problems the true functional relationship between a response variable Y , and the factors upon which it depends is not known. In many cases, especially in economics or force structure analyses, it cannot be exactly defined since neither all of the factors nor their inter-relationships which affect the response can be determined. In most cases, however, it is possible to isolate factors which might affect the response variable and, thereby, provide valuable insight into the mechanics of the problem at hand. After the selection of factors, the coefficients (b_1 , b_2 , b_3 , b_4) in the relationship must be defined. These coefficients show a relationship between the response variable Y and the factors. They potentially show the contribution of each factor to the value of Y . As with the factors, the true coefficients cannot be determined except in the physical sciences where exact laws govern the relationships. But, acceptable estimates of their true values can be determined by use of multiple regression analysis.

Consider the simple problem depicted in Figure 1 and allow the B1 PA to be variable also. The normal approach would be to leave Figure 1 as is and to conduct another one-dimensional analysis, varying the B1 PA while holding the number of B2s constant. Other data points could be collected by varying the B1 PA over the same range, but changing the number of B2s to a different value for each case. This approach, which requires many calculations, provides point data and can be illustrated as in Figure 3.

Notice that Figure 3 does provide more information. However, these data do not fully develop the relationship between DE, the B1 PA, and the number of B2s. Further, the plotted data does not provide an explicit measure of the contribution of each factor to the DE, nor does it explicitly depict potential inter-relationships between factors.

The shaded area of Figure 4 depicts the appropriate response surface for this problem. It extends the limited presentation of Figure 3 to a more fully developed relationship. The example point designated by the X is on the response surface. The multidimensional presentation



NUMBER OF B2s

FIG. 3

Example of Current

Method of Sensitivity Analysis

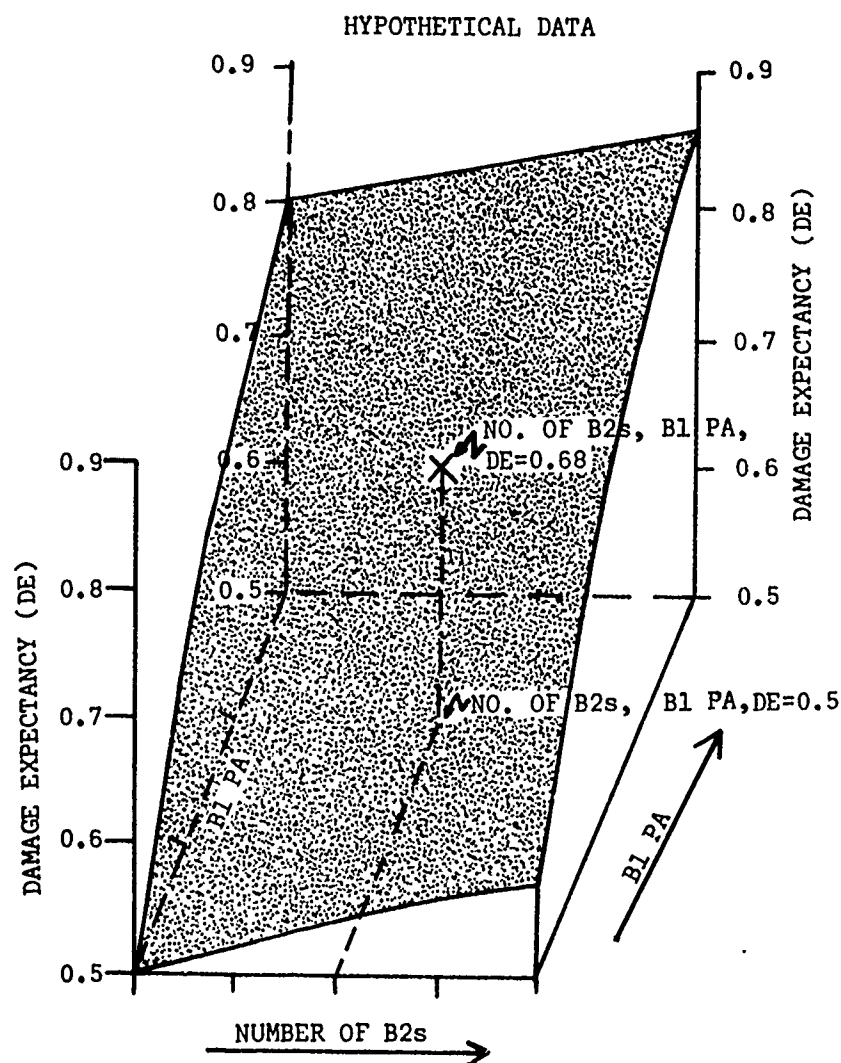


FIG. 4

Example Response Surface of DE Versus
B1 PA and No. of B2s

shows how the response variable for our problem, DE, varies with the various combinations of the factors, B1 PA and the number of B2s, that we have assumed in our example problem. The information depicted in Figure 4 can also be expressed in a functional mathematical form. The equation to this response surface is:

$$DE = b_0 + b_1 \times B1 \text{ PA} + b_2 \times \text{No. of B2s} + b_3 \times (B1 \text{ PA})^2 + b_4 \times (\text{No. of B2s})^2 \quad (2)$$

Again, if the coefficients (b_0 , b_1 , b_2 , b_3 and b_4) can be determined the equation can be used in three valuable ways. First, "what if" questions can be answered economically in real time without the requirement to run further analysis cases using a complicated computer program. Secondly, the contribution to DE per unit of each factor is readily available. And thirdly, ratios may be taken of the coefficients to provide a rough comparison of the contribution of one system versus another system. Further analysis can be conducted using the coefficients as a relative measure of worth to overall DE of improving a characteristic such as weapon system accuracy (CEP) or PA.

The three-dimensional example case can be extended to include a reasonable number of other factors. These factors can be weapon system characteristics, number of a particular type of weapon system or any other descriptive data that would be of interest in a particular problem. The descriptive equation (similar to equation 2) would be expanded to account for these other factors. However, to obtain the true surfaces that have been given in the simple example above would require a very large number of computer runs and be impractical for most applied analyses.

A methodology does exist that provides the capability to estimate the true response surface using a limited number of runs. The primary purpose of this methodology is to estimate the coefficients of an assumed mathematical equation and, thereby, define the shape of a response surface of outputs of a mathematical programming algorithm or other formulated decision problems. Obviously, if the equation includes more than two factors, it would be impossible to show a picture of the full relationships similar to the response surfaces of our example problem since it would be more than three-dimensions. However, it

is possible to show sections of this multidimensional surface. And the equation for the response surface still allows various analyses using the coefficients.

3. THE METHODOLOGY

The methodology involves the combination of Mathematical Programming techniques with Response Surface Methodology and Statistical Experimental Design [7]. Besides providing a capability to conduct multidimensional sensitivity analyses, the methodology provides a measure of effectiveness technique which is new to applied operations research technology and force structure analyses. The methodology does not provide the answers to force structure problems. It provides information which can be used in conjunction with operational, political and economical aspects to help point the direction.

Basically, the methodology uses Statistical Experimental Design to optimally select a limited number of combinations of the input parameter values to be evaluated in the analysis process. Mathematical Programming techniques such as linear or quadratic programming can then be used to find the optimum value of the response variable, for example DE, for each of the selected combinations of input parameters. Finally, the coefficients of the mathematical expression that defines the response surface can be determined by a multiple-linear regression technique. Once the equation is completely developed, the value of the response variable can be determined with good accuracy for various combinations of factor values, not just the ones used to develop the equation. Selected combinations of factors must be within the ranges of the parameters used to develop the equation since this type of equation can be very untrustworthy when extrapolated. The equation should be regarded only as a good approximation of the true response surface over the range of values of interest.

4. EXAMPLE PROBLEM

To demonstrate the concept and procedure consider a hypothetical strategic force of four weapon systems. Any one or all of the systems can be expanded in number. Our objective is to highlight the relative value of each system to the total force. The four systems and their range of numbers are given in Table 1.

TABLE 1
TYPES AND LEVELS OF WEAPON SYSTEMS²

<u>Weapon Type (Factors)</u>	<u>Min No.</u>	<u>Max No.</u>
ICBM - type 1 (M1)	300	450
ICBM - type 2 (M2)	400	550
Bomber - type 1 (B1)	150	250
Submarine - type 1 (S1)	10	28

The weapon system characteristics of each system such as CEP, WSR, PLS, PA, Yield, and number of weapons or warheads are given and remain constant for this example. The input variables (factors) in the example problem are the numbers of each type weapon system. Assume that Damage Expectancy (DE) is the response variable of interest to the decision maker.

5. THE PROCESS

The first step is to determine the particular combinations of factor levels using experimental design. There are a large number of possible combinations of the factors if we let each factor vary by one at a time. The total number of possible combinations is equal to:

$$(151)^2 (101)^1 (19)^1 = 43,755,119.$$

If all of these combinations could be run and the DE for each case plotted in five dimensions, the result would be the true model response surface. Obviously, it is desirable to find some smaller number of the combinations which will estimate the true surface as accurately as possible. The set of combinations selected for evaluation is called the experimental design.

2. These numbers have been arbitrarily selected for the purpose of illustration.

In selecting the experimental design, it is necessary first to postulate a form of the mathematical equation which will be used to approximate the surface. For this case a second order equation is assumed³:

$$\begin{aligned} DE = & b_0 + b_1 B1 + b_2 M2 + b_3 M1 + b_4 S1 + b_5 B1^2 \\ & + b_6 M1^2 + b_7 M1^2 + b_8 S1^2 + b_9 (B1 \times M2) \\ & + b_{10} (B1 \times M2) + b_{11} (B1 \times S1) + b_{12} (M1 \times M2) \\ & + b_{13} (M2 \times S1) + b_{14} (M1 \times S1) \end{aligned} \quad (3)$$

where, as before, B1 = number of Bombers-type 1. M2 = number of ICBM-type 2. M1 = number of ICBM-type 1, and S1 = number of Submarines-type 1. The problem in experimental design is to select a specific number of combinations from all possible which best estimates the coefficients $b_0, b_1, b_2, \dots, b_{13}, b_{14}$; of equation (3). To estimate coefficients for a second order equation, it is necessary to identify a minimum of three levels or values for each of the factors. For this problem, the three levels chosen are the values at the upper and lower end of the range for each input parameter plus a point chosen so that it is the mid-point between the end points. The third point is chosen this way in order that the values may be coded for easy calculation and to produce coefficients which are non-correlated. The points for this example problem and their coded values are given in Table 2.

There are a limited number of experimental designs available for second-order equations [4] [8]. But there are no practical optimal designs available in the research literature for response surfaces where the variance of the response is equal to zero [5] [6]. The DE derived from a mathematical programming problem has no variance because whenever the same factor values are input into the DE optimi-

3. This is usually sufficient in almost all real world problems. The reason is that a second order equation usually gives a good approximation over small areas of real world surfaces.

zation model, it always gives the same answer. All practical designs in the literature are for problems in which the response has a variance. [3]. However, as will be seen, these designs provide very satisfactory results for response surfaces which have no variance but where the error in the fitted surface is due to bias only. The design chosen for this example problem is given in Table 3. It is a three level design without the repetitive center design points [2].

Table 2
Factor Levels

<u>Weapon Types (Factors)</u>	<u>Levels (Codes)</u>			
B1	150 (-1)	200 (0)	250 (+1)	
M2	400 (-1)	475 (0)	550 (+1)	
M1	300 (-1)	375 (0)	450 (+1)	
S1	10 (-1)	19 (0)	28 (+1)	

Following is an example of how the values are coded:

$$\text{Coded B1} = -1 = \frac{150 - 200}{50}$$

$$\text{Coded B1} = 0 = \frac{200 - 200}{50}$$

$$\text{Coded B1} = +1 = \frac{250 - 200}{50}$$

The 25 runs in the design will provide data for estimating the coefficients, the b's, in equation (3).

Having decided upon an experimental design, the number of combinations and the value of the factor levels for each combination, the next step is to make 25 force allocation runs using linear programming. One run is made for each of the identified combinations in order to get the optimum DE for each set of input factor values.

Table 3
Experimental Design

Run#	<u>Non-Coded</u>				<u>Coded</u>			
	M2	M1	B1	S1	M2	M1	B1	S1
1	550	450	200	19	1	1	0	0
2	550	300	200	19	1	-1	0	0
3	400	450	200	19	-1	1	0	0
4	400	300	200	19	-1	-1	0	0
5	475	375	250	28	0	0	1	1
6	475	375	250	10	0	0	1	-1
7	475	375	150	28	0	0	-1	1
8	475	375	150	10	0	0	-1	-1
9	550	375	200	28	1	0	0	1
10	550	375	200	10	1	0	0	-1
11	400	375	200	28	-1	0	0	1
12	400	375	200	10	-1	0	0	-1
13	475	450	250	19	0	1	1	0
14	475	450	150	19	0	1	-1	0
15	475	300	250	19	0	-1	1	0
16	475	300	150	19	0	-1	-1	0
17	550	375	250	19	1	0	1	0
18	550	375	150	19	1	0	-1	0
19	400	375	250	19	-1	0	1	0

Table 3 - Continued

<u>Run#</u>	<u>Non-Coded</u>				<u>Coded</u>			
	<u>M2</u>	<u>M1</u>	<u>B1</u>	<u>S1</u>	<u>M2</u>	<u>M1</u>	<u>B1</u>	<u>S1</u>
20	400	375	150	19	-1	0	-1	0
21	475	450	200	28	0	1	0	1
22	475	450	200	10	0	1	0	-1
23	475	300	200	28	0	-1	0	1
24	475	300	200	10	0	-1	0	-1
25	475	375	200	19	0	0	0	0

The final step in deriving the equation is to use the results of the 25 cases as input to a multiple-linear regression program. The coded levels of each of the weapon systems or factors are input to the program as independent variables and the DE is input as the response or dependent variable. Using multiple-linear regression theory the coefficients of equation (3) are determined and those which are significant are retained in the equation. The major tests of how well the second-order equation approximates the true surface are the percentage of the total variation in the DE values that is explained by the factors in the equation, and the magnitude of the prediction errors when the resultant equation is used to predict the 25 values input to the regression model. Equation (4) is the equation for the example problem using the coded values in Table 3.

$$\begin{aligned}
 \text{DE} = & \overset{(b_0)}{0.59363} + \overset{(b_1)}{0.036} B1 + \overset{(b_2)}{0.01072} M2 + \overset{(b_3)}{0.00397} M1 \\
 & + \overset{(b_4)}{0.02976} S1 - \overset{(b_5)}{0.00796} B1^2
 \end{aligned}
 \tag{4}$$

Coefficients denoted by b_6 through b_{14} in equation (3) did not pass the significance tests. To convert the coded equation (4) to a relationship in which the actual numbers of the weapon systems can be used, it is necessary to substitute the ratios used to develop the (-1, 0, +1) values back into equation (4). Substituting

$$\begin{array}{ll} \frac{B1 - 200}{50} \text{ for } B1 & \frac{M2 - 475}{75} \text{ for } M2 \\ \frac{M1 - 375}{75} \text{ for } M1 & \frac{S1 - 19}{9} \text{ for } S1 \end{array}$$

into equation (4) gives:

$$\begin{aligned} DE = & 0.1717 + 0.001994 B1 - 0.000003184 B1^2 + 0.000143 M2 \\ & + 0.000053 M1 + 0.00307 S1 \end{aligned} \quad (5)$$

This final form is the mathematical functional relationship which is the desired response surface in equation form. It shows the relationships of each factor (weapon system) to the response (DE) and provides a method of evaluating the value of any factor relative to the contribution of another factor. The b_0 value, 0.1717, is the contribution to DE from all other weapon systems not included as factors in the analysis.

The accuracy of this equation is demonstrated by the data in Table 4 which contains the actual results of force structure allocations obtained by the linear programming runs compared to those values predicted by the derived relationship shown in equation (5).

For the purpose of analyses, the accuracy of the DE predicted by equation (5) is excellent.

6. EXAMPLE ANALYSIS

Equation (5) can now be used to provide information to help answer various questions. Figures 5 through 8 give examples of the use of the response surface formulation. The presentation in Figures 5, 6, and 7 use the coefficients for each factor to illuminate the relationships between the factors and the response variable and the relative

value of each system as compared to other systems. There are no further calculations required to obtain the information in these figures.

Table 4
Actual DE Versus Predicted DE Values Using Hypothetical
Weapon System Data

<u>Weapon Levels</u>				<u>Actual DE</u>	<u>Predicted DE</u>	<u>Errors</u>
B1	M2	M1	S1			
160	420	400	14	0.5398	0.5335	0.0063
180	420	400	14	0.5526	0.5576	0.0050
220	420	400	14	0.5775	0.5805	0.0030
230	500	450	24	0.6344	0.6309	0.0035

Figure 5 shows the DE improvement for each weapon system in terms of the number of carriers. Notice in equation (5) that the only weapon system having a squared term associated with it is the B1 weapon system. The other weapon systems are included in the equation solely in linear terms. Figure 5 depicts this by showing how much a single carrier improves the overall total force DE. For example, the DE improvement per M2 is 0.000143 regardless whether or not there are 400 or 550 M2 carriers. Because there is a squared term associated with the B1 weapon system in the equation, the average contribution per B1 to the total force DE is not constant. The contribution changes as a function of the number of B1s as is shown in Figure 5. The average contribution per B1 is 0.00153 for 150 B1s and 0.0012 for 250 B1s.

Figure 6 shows the same information as in Figure 5 except that the data is for the contribution to DE per warhead. The data shows a constant return per warhead for the M1, M2, and S1 warheads. Again, there is a diminishing marginal return concept associated with the B1 system. This says that the value of a B1 warhead decreases as the number of B1s increase for a given threat target base. Figures 5 and 6 show the same information, but presented in two different ways.

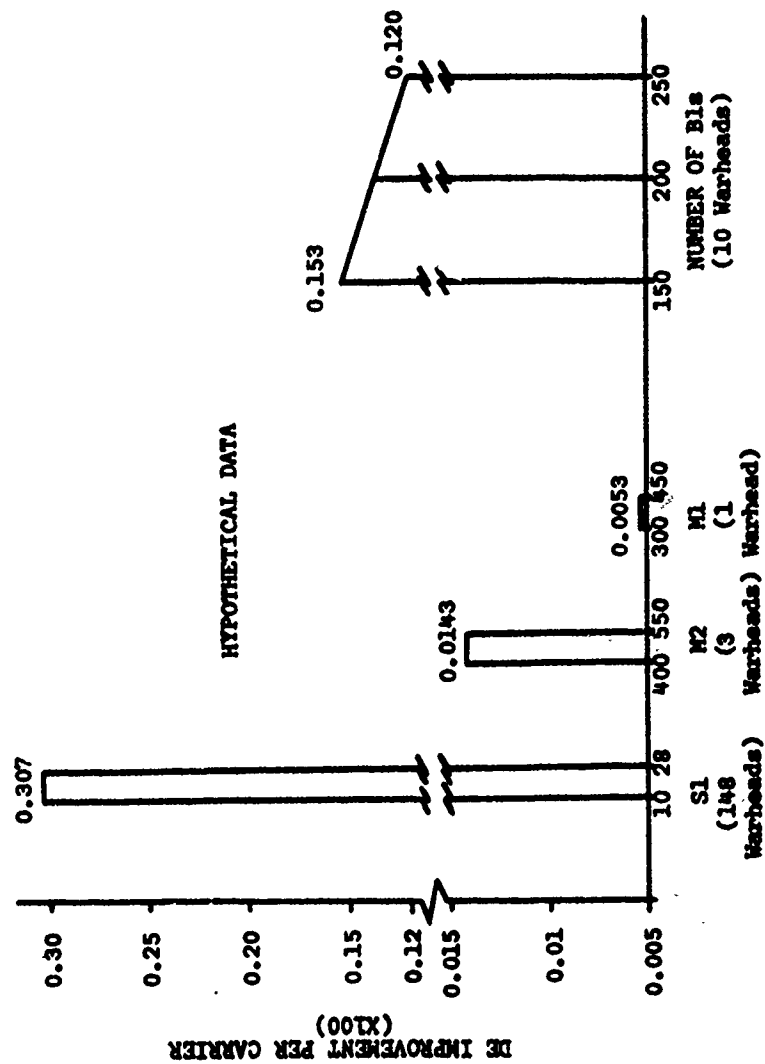


FIG. 5

DE Improvement for Each Weapon System in Terms of Carriers

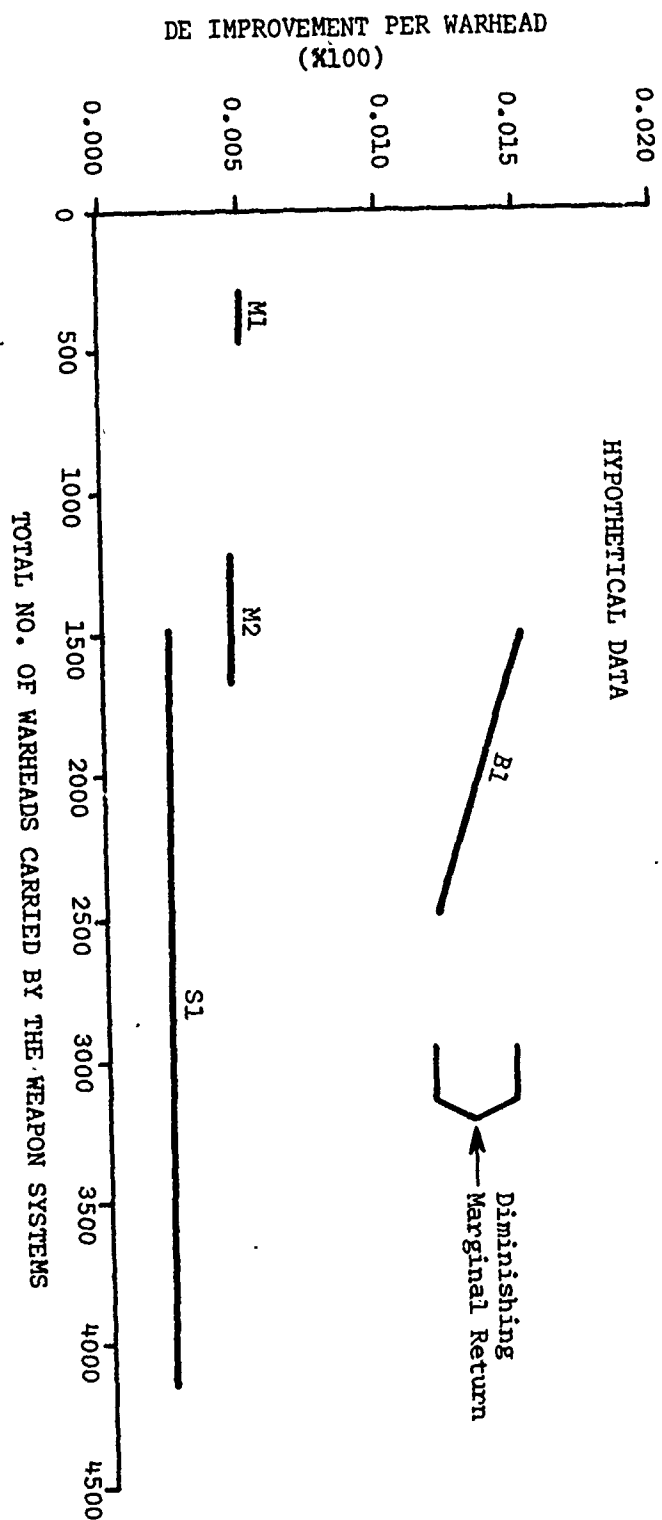


FIG. 6

DE Improvement for Each Weapon System in Terms of the Number of Warheads

Another capability that is available by use of the response surface equation is the comparisons of the overall contribution of each weapon system. First, multiplying the coefficients times the total number of weapons within the range gives the contribution of that particular weapon to the overall damage expectancy. Secondly, the coefficients themselves give the contribution per unit of a particular weapon system (factor) to DE. Because the coefficients of equation (5) are non-correlated, a ratio can be obtained using them which gives the relative contribution of one system as compared to another. This is the information that is depicted in Figure 7. For example, using 150 B1s in the force as the basis, one B1 within the constraints of the problem is equivalent to 1/2 of a S1 with 148 warheads. One B1 is also equivalent to 10 1/2 M2s with 3 warheads. When 250 B1s are in the force, 1 B1 is equivalent to 3/10 of a S1 and 8 M2s. It is important to note at this point that this value scheme is based solely upon Damage Expectancy and if a different decision criterion was selected, e.g., the capability to hit targets in a timely manner, the M1 or M2 may be worth more than a B1.

Figure 8 presents four different factor values and their relationships. Using equation (5), it is very easy to develop constant DE contours for various combinations of weapons systems. In Figure 8 the number of B1s, the number of S1s, and the number of M2s are varied over the previously selected ranges. The number of M1s was arbitrarily fixed at 450, although this was not necessary. Figure 8 shows contours for a 0.55 DE through a 0.65 DE for various numbers of M2s. This figure shows, for example, that any combination of B1s, M2s and S1s within this band will give a total force DE of 0.55, keeping the number of M1s equal to 450. Also consider the first vertical area shown, 14 to 18 S1s. This adds a fifth element to the data already depicted in the form of a constraint such as might be imposed under a strategic arms limitations (SAL) agreement. For example, perhaps particular terms in SAL indicate that the number of S1s should be between 14 and 18. The vertical area gives all the possible combinations of forces that are possible for a selected damage expectancy. Likewise, the second vertical area denoted in the range from 20 to 24 S1s gives the same information, but at a different level.

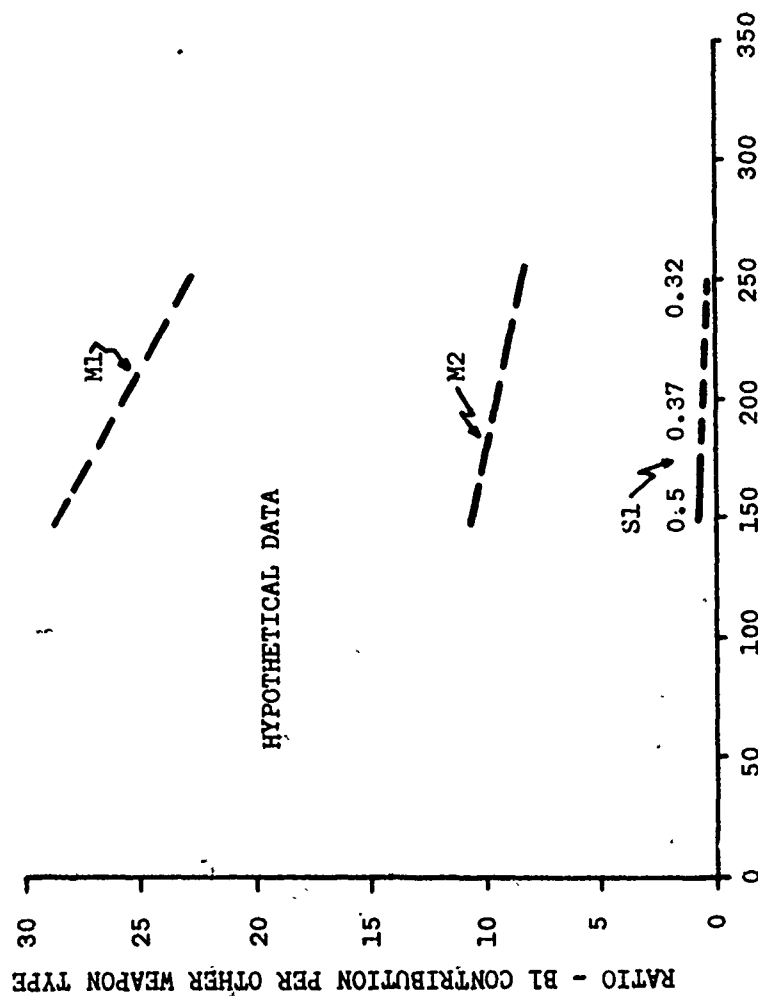


FIG. 7
Ratio of the Contribution to DE of a B1 to One Carrier of Other
Weapon Types

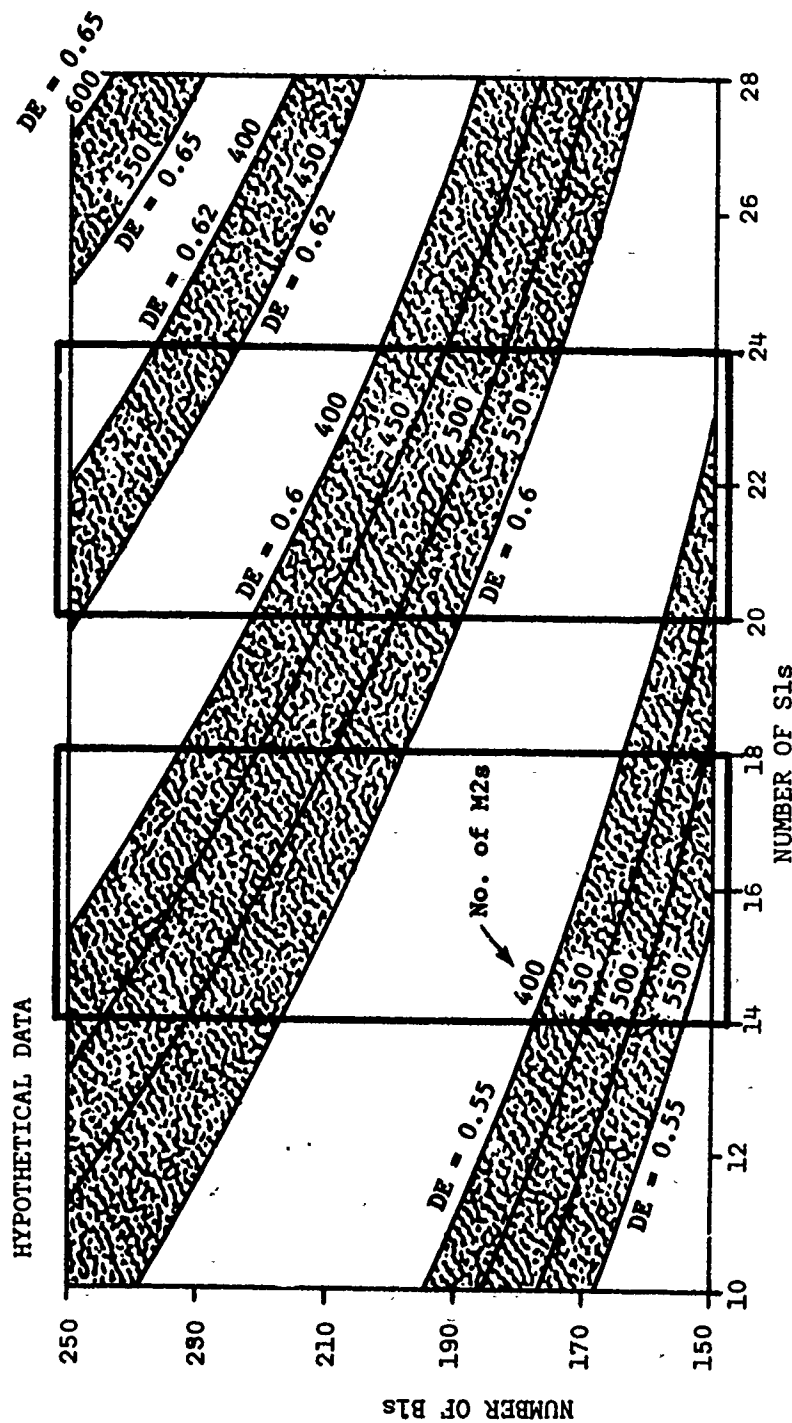


FIG. 8
Constant DE Contours for Various Combinations of SIs, Bls, and M2s.
Number of MIs is Constant at 450.

Many other types of information can be derived easily and economically from the equation of the response surface. For example, the number of M2s and the number of B1s could be changed at the same time in an equal or unequal ratio. Constant damage expectancy plots for different constant ratios for each of the weapon systems could be generated easily. Any combination of the factors could be changed to other values within the original ranges of the factor values. All without making additional computer runs.

7. SUMMARY AND CONCLUSIONS

The multiple dimensional parametric analysis technology using response surface concepts provides an expanded capability for conducting a more valuable analysis in a complex environment. It provides a picture of what is happening within the model being used for the study. It also provides insight into the relationship among the factors under study. "What if" analyses can be conducted economically and in real time without the necessity to obtain new computer outputs.

The contribution to the selected measure of effectiveness (MOE) per unit of each factor and ratio comparisons of the different coefficients provide a new measure of effectiveness tool for comparing the worth and capability of one system vs another within the context of the model used in an analysis. However, it is important that an analyst know the problem well in order to properly interpret results. For example, if the number of B2s was a planning factor under study and yield, CEP, and reliability of its weapons were also factors under study, the data would be correlated. The contribution of per unit of yield, for example, to the overall force DE would be a function of the number of B2s. This type of insight is needed in order to obtain a maximum contribution from the methodology.

As far as the mechanics of using the methodology are concerned, they are very simple. First, define the model; secondly, define the experimental design; thirdly, obtain the values of the MOE for each run; fourthly, fit a response surface to the data using multiple-linear regression techniques. The methodology can be used with any decision model, linear or non-linear in form.

REFERENCES

- [1] Box, G.E.P., THE EXPLORATION AND EXPLOITATION OF RESPONSE SURFACES, Biometrics, Vol 10, PP. 16-60, March 1954.
- [2] Box, G.E.P. and D.W. Behnken, SOME NEW THREE LEVEL DESIGNS FOR THE STUDY OF QUANTITATIVE VARIABLES, Technometrics, Vol 2, PP. 455-475, November 1960.
- [3] Box, G.E.P. and N.R. Draper, A BASIS FOR THE SELECTION OF A RESPONSE SURFACE DESIGN, J. Amer. Stat. Assoc., Vol 54, PP. 622-654, September 1959.
- [4] Draper, N.R. and D.M. Stoneman, RESPONSE SURFACE DESIGNS FOR FACTORS AT TWO AND THREE LEVELS AND AT TWO AND FOUR LEVELS, Technometrics, Vol 10, No. 1, PP. 177-192, February 1968.
- [5] Karson, M.J., DESIGN CRITERION FOR MINIMUM BIAS ESTIMATION OF RESPONSE SURFACES, J. Amer. Stat. Assoc., Vol 65, No. 332, December 1970.
- [6] Karson, M.J., Manson, A.R. and R.J. Hader, MINIMUM BIAS ESTIMATION AND EXPERIMENTAL DESIGN FOR RESPONSE SURFACES, Technometrics, Vol 11, No. 3, August 1969.
- [7] Smith, P.W., A METHODOLOGY FOR DEVELOPING AND ANALYZING THE OPTIMAL RESPONSE OF AN ECONOMIC CRITERION TO SIMULTANEOUSLY INDUCED MULTIPLE EVENTS USING INPUT-OUTPUT, MATHEMATICAL PROGRAMMING, AND RESPONSE SURFACE METHODOLOGIES, Unpublished Ph.D. Dissertation, Univ. of Alabama, July 1975.
- [8] Webb, S.R., SMALL INCOMPLETE FACTORIAL EXPERIMENT DESIGNS FOR TWO- AND THREE-LEVEL FACTORS, Technometrics, Vol 13, No. 2, PP. 243-256, May 1971.